Combining 3D Convolutional Neural Networks with Transfer Learning by Supervised Pre-Training for Facial Micro-Expression Recognition

Ruicong ZHI†, ‡, Hairui XU†, ‡, Ming WAN†, ‡, Nonmembers, and Tingting LI†, ‡, Member

SUMMARY
Facial micro-expression is momentary and subtle facial reactions, and it is still challenging to automatically recognize facial micro-expression with high accuracy in practical applications. Extracting spatiotemporal features from facial image sequences is essential for facial micro-expression recognition. In this paper, we employed 3D Convolutional Neural Networks (3D-CNNs) for self-learning feature extraction to represent facial micro-expression effectively, since the 3D-CNNs could well extract the spatiotemporal features from facial image sequences. Moreover, transfer learning was utilized to deal with the problem of insufficient samples in the facial micro-expression database. We primarily pre-trained the 3D-CNNs on normal facial expression database Oulu-CASIA by supervised learning, then the pre-trained model was effectively transferred to the target domain, which was the facial micro-expression recognition task. The proposed method was evaluated on two available facial micro-expression datasets, i.e. CASME II and SMIC-HS. We obtained the overall accuracy of 97.6% on CASME II, and 97.4% on SMIC, which were 3.4% and 1.6% higher than the 3D-CNNs model without transfer learning, respectively. And the experimental results demonstrated that our method achieved superior performance compared to state-of-the-art methods.

key words: facial micro-expression, 3D convolutional neural networks, transfer learning, spatiotemporal features

1. Introduction

Facial expression recognition is a popular research topic that has been extensively studied in recent years. Although the research on the facial expression recognition began in 1970s [23], the study of the facial micro-expression is still rare mainly because humans tend to recognize ordinary facial expression which presents rich changes, rather than subtle facial expression.

Facial micro-expression is momentary involuntary facial expression which usually occurs in high-stake situations, when people undergo but do not intend to express. Facial micro-expression lasts an extremely short period which does not exceed 1/2 second [25], [31]. There are obviously differences between facial micro-expressions and macro-expressions in the aspect of short duration and occurrence [28].

There are numerous potential applications for facial micro-expression recognition recently, and analyzing facial micro-expression has been a relative new subject in clinical diagnosis, education, business, and criminal investigation [2], [3], [22]. For example, facial micro-expression could be used to help doctors identifying the patient’s true feelings. In addition, facial micro-expression can serve as an important cue to detect lying in criminal interrogation since it unconsciously displays when he/she tries to hide his/her emotion. Teachers can read the psychological changes of students through facial micro-expressions to adjust the classroom atmosphere, stimulate students’ interests in learning, and communicate with students better. In business, the introduction and application of facial micro-expression have important reference value and a certain significance in the full understanding of the client’s inner world and more accurately grasp the true demands of clients [15]. Since the recognition of facial micro-expression by humans is challenging, it is necessary to study automatic facial micro-expression recognition technology to improve the accuracy and efficiency of the task.

1.1 Related Work

Automatic facial micro-expression recognition is a challenging task, and has attracted increasing interests in computer vision. Generally, there are two main types of facial representation features, such as geometric and appearance features [4]. Geometric-based methods calculate the distance between facial landmarks in the facial image to form a feature vector which represents the shape and locations of facial components. While the appearance-based methods apply the descriptors, to describe the skin texture of the face, caused by expression [7].

Geometric-based method is a simple model method, which uses fiducial landmarks or template to describe the shape structure of human face. Active Appearance Model (AAM) and Active Shape Model (ASM) are commonly applied geometric-based methods. Park et al. [20] applied the AAM to extract 70 facial landmarks from image sequence, and utilized three facial landmarks (two eyes and nose tip) to align the image sequence. Then, the motion vectors of 27 feature points were estimated by the feature point tracking method. Finally, the subtle facial expressions were transformed to corresponding exaggerated ones by magnifying

Manuscript received May 3, 2018.
Manuscript revised December 19, 2018.
Manuscript published January 29, 2019.
†The authors are with the School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, P.R. China.
‡The authors are with Beijing Key Laboratory of Knowledge Engineering for Materials Science, Beijing 100083, P.R. China.
a) E-mail: zhrc_research@126.com
DOI: 10.1587/transinf.2018EDP7153
the motion vectors of the 27 facial landmarks. Michael et al. [17] proposed an automatic and covert method to detect deception in interrogation interviews by using motion profiles extracted from the motion of hands, and micro-expression with skin blog analysis and ASM.

In the midst of appearance-based methods, Local Binary Pattern (LBP) is regarded as a well-applied texture descriptor due to its simplicity and efficiency [24]. In order to extract the appearance and motion features across spatiotemporal dimensions, the traditional LBP descriptor is extended to LBP-TOP (LBP from three orthogonal planes), which is the 3D variance of LBP. As a result, many researchers are interested in the potential ability of LBP-TOP for facial micro-expression analysis [22], [29], [36], by considering its dynamic characteristics and efficient computation. Li et al. [12] employed LBP-TOP on their proposed facial micro-expression database (SMIC) which was established by University of Oulu and released a baseline performance of up to 48.78% accuracy for three facial micro-expression classes. In the approach of Li et al., temporal interpolation model (TIM) [22] was adopted to extract statistically stable features of the facial micro-expression video clips with sufficient number of frames.

In recent years, deep learning is increasingly showing its potential in image analysis, such as image classification, object detection and facial expression recognition. However, the studies of deep learning applied to facial micro-expression recognition are limited, and it is mainly due to insufficient labeled samples in the available facial micro-expression database.

Patel et al. [21] first applied deep learning for facial micro-expression recognition and used feature selection to improve the performance, and the reported accuracy on SMIC and CASME II were 53.6% and 47.3%, respectively. Kim et al. [9] proposed a new feature representation which was learned with expression-states by using convolutional neural networks (CNN) and long short-term memory (LSTM) recurrent neural networks and achieved 60.98% accuracy on CASME II.

1.2 Contribution

To the best of our knowledge, in computer vision, the ordinary facial expression recognition has successfully implemented and achieved accuracies of over 90% on several facial expression databases [14]. However, researches on facial micro-expression are rare due to two reasons. One is the lack of well-established database [30], and the other is the challenge to capture facial motion information in facial micro-expression image sequences. In contrast to facial macro-expressions, short duration and subtle facial changes are the most important characteristics of facial micro-expression [32].

There is a growing evidence of psychological research that temporal dynamics of facial behavior is a critical factor for the interpretation of the observed behavior. It is even important to capture the dynamics of temporal facial micro-expression along spatiotemporal dimensions, as they could reflect the extremely minor changes in facial micro-expression. One contribution of our study is employing 3D Convolutional Neural Networks (3D-CNNs) to learn spatiotemporal features from facial micro-expression image sequences. 3D-CNNs can model appearance and motion simultaneously for video data, and they are successfully applied in action recognition [8]. It is helpful to utilize them to perform high-quality facial micro-expression recognition.

Moreover, in order to solve the problem of insufficient samples in facial micro-expression database, we employ a transfer learning method through pre-training the deep learning model by leveraging the existing labeled data of related domain. Then the knowledge learned from source data is applied to the task of interest [34], which is the other contribution of this study.

In summary, we exploit 3D-CNNs to learn spatiotemporal features from facial micro-expression image sequences, and spatial augmentation is utilized on each sequence to reduce potential overfitting and improve classification accuracy. There are two stages in our system. In the first stage, the 3D-CNNs are conducted on the traditional facial expression database (Oulu-CASIA) by supervised learning. Then, the knowledge and pre-trained model learned from source domain are transferred to the target domain in the second stage. The final 3D-CNNs model is performed on two available published facial micro-expression datasets, i.e. CASME II and SMIC. Experimental results demonstrate that our method can achieve outperforming performance than state-of-the-art.

2. Methodology

2.1 Overview

The overall strategy of our facial micro-expression recognition system is depicted in Fig. 1. In preprocessing procedure, 3D spline interpolation [35] is utilized to normalize the length of the input facial image sequences. The clips with too many frames are down-sampled along the time axis and the length of every image sequence is normalized to 10 frames. Moreover, to avoid overfitting, data augmentation is applied in target task, and each image sequence is flipped horizontally, and the original dataset is extended by seven times. Firstly, we pre-train the 3D-CNNs model on facial expression database Oulu-CASIA by supervised learning and store the knowledge learned from source domain. Then, the pre-trained model is transferred to the target domain (facial micro-expression). Finally, we extract the feature vectors from the last fully-connected layer as feature representation of each facial micro-expression sequence, these feature vectors can describe the facial texture and motion characteristics very well, then these feature vectors are fed into a Linear SVM to output the emotion class.
2.2 Three-Dimensional Convolutional Neural Networks (3D-CNNs)

The 3D-CNNs show a good performance in video-based analysis due to its ability to learn spatiotemporal features. In 2D CNNs, convolution and pooling operations are applied on 2D images or feature maps spatially without taking consideration of motion information [26]. Therefore, 3D convolution can preserve temporal information of the input signals effectively compared to 2D convolution.

The 3D convolution operations are performed by convolving the 3D filters on the frame cube and sharing the weights along the time axis [8]. Generally, the value at position \((x, y, z)\) in the \(j\) th output feature map in the \(i\) th layer is calculated as follows:

\[
 v_{ij}^{xyz} = \sigma\left( \sum_{m} \sum_{p=0}^{P-1} \sum_{q=0}^{Q-1} \sum_{r=0}^{R-1} w_{ijm}^{pqr}(x+p)(y+q)(z+r) + b_{ij} \right), \tag{1}
\]

where \(w_{ijm}^{pqr}\) is the \((p, q, r)\) th value of the filter connected to the \(m\) th output feature map in the previous layer. \(P_i\) and \(Q_i\) are the height and width of the 3D filter, respectively, \(R_i\) is the depth of the 3D filter along the temporal dimension, and the nonlinear activation function \(\sigma(x) = \max(0, x)\) is utilized in our model, named Rectified Linear Units (ReLU) [18].

Figure 2 illustrates the 3D convolutional kernel (depth=3) applied on the cube formed by stacking multiple frames or feature maps. These connections are color-based so that the same color means that the frames share the same weights. Particularly, a 3D convolutional kernel can only extract one type of feature from the frame cube, since the kernel weights are shared across the entire cube in the temp-
poral dimension. When the 3D convolutional kernel acts on the input cube, the process can be treated as three 2D convolutional kernel acting on the contiguous three frames, respectively. Finally, the convolution results are integrated and output feature maps are obtained. By this construction, the 3D convolution operations can preserve the temporal information of the input signals and extract the spatiotemporal features from contiguous frames. The same construction is applied for 3D pooling.

2.3 Transfer Learning by Supervised Learning

Transfer learning aims at extracting the knowledge from source domain and utilizes the learned knowledge to obtain a better model in the target domain \[19\]. There are different ways to transfer learned knowledge from source task to target task. Due to the deficit of samples for facial micro-expression recognition, it is essential to pre-train the model on related domain. In addition, this is an effective way to avoid diffusion of gradients in deep networks without adequate training samples. Usually, networks layers are initialized randomly, which probably leads to the network getting stuck in local optima and could not learn meaningful features. To deal with the problem of insufficient samples in the database, the normal facial expression database Oulu-CASIA NIR is utilized in pre-training stage as the source data.

We select five-class (anger, disgust, happiness, sadness and surprise) labeled samples from Oulu-CASIA NIR to pre-train our 3D-CNNs model for facial micro-expression recognition on CASME II and three-class (fear, happiness and surprise) labeled data for SMIC, since the class number of facial micro-expression is different on the two databases. As illustrated in Fig. 1, the model is pre-trained on the source task (Oulu-CASIA NIR) with five-class or three-class labeled data firstly. Then the parameters of the overall network are stored and transferred to the target task (facial micro-expression recognition on CASME II and SMIC-HS). Finally, we extract the feature vectors from the last fully-connected layer to train a Linear SVM for facial micro-expression classification.

3. Implementation

3.1 Preprocessing

The length of facial micro-expression sequence in CASME II and SMIC database is different. 3D spline interpolation \[35\] was applied to normalize the length of input sequence to 10 frames. Figure 3 represented a original cropped sample of facial micro-expression sequence with

![Fig. 3](image-url) The original 64 frames of a sample sequence of CASME II.

![Fig. 4](image-url) The interpolated sequence for the same sample sequence of CASME II.
64 frames in CASME II, while Fig. 4 showed the interpolated sequence for the same sample sequence. It could be observed that the key frames with significant changes were preserved in the interpolated sequence. After the interpolation of the facial micro-expression sequences, the facial features were more obvious, and the feature extraction was simpler.

Data augmentation was applied to avoid overfitting. Each image sequence was flipped horizontally, and rotated according to angles in \{-7°, -5°, -3°, 3°, 5°, 7°\}, and the size of input image was normalized to 32 × 32 with gray scale. As a result, we obtained seven times more data combined with original image sequences.

### 3.2 3D-CNNs Architecture

In this section, we described the network settings of our 3D-CNNs model. The model consisted of two 3D convolutional layers and two pooling layers (each convolutional layer was followed by a pooling layer), two fully-connected layers, where the last fully connected layer was softmax output layer to predict facial micro-expression labels. We used 10 image frames with size of 32 × 32 as input of the 3D-CNNs model. All convolutional kernels were set to the size of 3 × 5 × 5, where 3 was the temporal depth and 5 × 5 was the size of the receptive field spatially. All the convolutional layers were used with VALID padding spatiotemporally which meant no padding and stride 1 × 1 × 1. For the pooling layer, it was max pooling with kernel size 2 × 2 × 2 and stride 1 × 2 × 2, which had stride 1 in temporal dimension in order to preserve the temporal signal, and the SAME padding was used for pooling layer which meant zero padding was introduced to make the shapes match as needed.

Details of the network architecture were given in Table 1. We trained the 3D-CNNs using mini-batches of 150 samples, with base learning rate of 0.001. The Adam optimizer was utilized for optimization and the dropout parameter was utilized for regularization.

| Layer type               | Filter size | Stride | Output size       | Padding |
|-------------------------|-------------|--------|-------------------|---------|
| Convolution             | 3 × 5 × 5   | 1 × 1 × 1 | 8 × 28 × 28 × 32  | VALID   |
| Relu                    | -           | -      | 8 × 28 × 28 × 32  | -       |
| Max-pool                | 2 × 2 × 2   | 1 × 2 × 2 | 8 × 14 × 14 × 32 | SAME    |
| Convolution             | 3 × 5 × 5   | 1 × 1 × 1 | 6 × 10 × 10 × 64  | VALID   |
| Relu                    | -           | -      | 6 × 10 × 10 × 64  | -       |
| Max-pool                | 2 × 2 × 2   | 1 × 2 × 2 | 6 × 5 × 5 × 64    | SAME    |
| Fully connected layer 1 | -           | -      | 256               | -       |
| Fully connected layer 2 | -           | -      | output labels     | -       |
| Dropout                 | -           | -      | -                 | 0.5     |

### 3.3 Pre-Training with Transfer Learning

The Oulu-CASIA NIR database [33] is a part of the Oulu-CASIA database recorded by a NIR camera. It includes 1440 image sequences with six basic facial expressions. Each image sequence starts at neutral and ends at a peak of the emotion. In our experiments, the dataset was split by a ratio of 80:20, and we used 80% data for model training and 20% data for observing the model state. The length of each facial image sequence in Oulu-CASIA NIR was normalized to 10 frames by employing 3D spline interpolation, and the input image was converted to 32 × 32. We initialized the parameters of our 3D-CNNs network from a normal distribution, i.e. N(μ = 0, σ² = 0.001). The exponentially learning rate was applied in our model as the training progress. The 3D-CNNs model was trained using mini-batches of 150 samples, with base learning rate of 0.0001 and decayed every 1000 steps with a base of 0.95.

### 4. Experiments and Discussion

To assess the performance of our approach, the experiments were conducted on CASME II [29] and SMIC [12] databases for facial micro-expression recognition. These databases were recorded in a controlled scenario and all the images were cropped. The architecture of our model was performed using the deep learning framework TensorFlow [27].

#### 4.1 Database Description

The CASME II database [29] consisted of 247 facial micro-expression video clips recorded by a 200fps camera from 26 participants, where the frames of onset and offset of each video clip were coded, as well as tagged with AUs and emotion. These clips were labeled with five kinds of expression: happiness (32 clips), surprise (25 clips), disgust (64 clips), repression (27 clips) and others (99 clips). A five-class baseline performance was released of up to 63.4% in the literature [29], in which LBP-TOP and SVM were employed for feature extraction and classification.

The SMIC database [12] included 164 spontaneous facial micro-expression video clips with the resolution 640 × 480 pixels from 16 participants. SMIC consisted of three sub-datasets, which were SMIC-HS (recorded by high speed camera), SMIC-VIS (recorded by normal visual camera), and SMIC-NIR (recorded by near-infrared camera). These clips were categorized into three types of expression: positive (51 clips), negative (70 clips) and surprise (43 clips). The three-class baseline performance was reported as 49.3% using LBP-TOP for feature description and SVM for classification.
4.2 Performance of Different Parameters

The performance of 3D-CNNs model was affected by different parameters. In this section, experiments were designed to evaluate the performance of our 3D-CNNs model with diverse image size and filter size.

4.2.1 Image Size

The various image sizes were set in the 3D-CNNs model to assess the performance. The image size was adjusted by taking consideration both computational efficiency and memory consumption. The initial size of the image was set to 36 × 42, and the image was scaled to a certain ratio. The performance of 3D-CNNs model with different image sizes on five subsets was listed in Table 2. The facial images were resized to rectangle and square size. When image size was small (e.g. 24 × 24), little useful information was remained leading to low recognition accuracy. Square facial images outperformed rectangle facial images, and the recognition accuracy increased when the height and width of the facial images are near equal. From the experimental results we may conclude that square images are more appropriate for facial micro-expression recognition, and the facial images were set to 32 × 32 in the subsequent section.

4.2.2 Filter Size

The 3D-CNNs could extract the motion information of video sequence effectively mainly depend on the 3D convolution kernel, i.e. the depth in the time dimension. In this section, the depth of the 3D convolution kernel was set from 1 to 5, and the results were shown in Table 3. It was illustrated that the difference of recognition accuracies was not obvious. Better performance was obtained when the depth was 2 and 3, which denoted the suitable filter size for spatio-temporal analysis. Du et al. [26] studied the filter size of 3D convolution through experiments and selected the 3D convolution kernel as 3 in their experiments. Therefore, by taking consideration of efficiency and computational complexity, the filter size of depth was set to 3 in the subsequent analysis.

4.3 The Effect of Pre-Training

Experiments were conducted to explore the effect of the pre-training, in which the pre-trained model was saved every 100 steps (from 0 to 600 steps). As a result, we obtained 7 pre-trained models on Oulu-CASIA NIR. In addition, the performances of three classifiers (Softmax, Linear SVM and Random Forest (RF)) were evaluated in the experiments, the estimators in the random forest was set to 30.

Figure 5 demonstrated the results of the pre-training experiments. Each data point in the line graph corresponded...
the facial movement information was not obvious for facial micro-expressions.

4.4 Experimental Results

For the two databases, we firstly applied 3D spline interpolation to normalize the length of the input sequences, and then spatial augmentation was utilized on each sequence. The 3D-CNNs model was pre-trained on Oulu-CASIA NIR and then transferred to target task for evaluation. To be more persuasive, the five-fold cross validation was employed in the experiments, that is, the entire dataset was divided into five subsets, and one of the five subsets was treated as testing set while the remaining four subsets were used for training. Finally, the Linear SVM was employed for classification. Moreover, the leave-one subject-out validation was also conducted for better comparison to related literatures.

4.4.1 CASME II

The confusion matrix of the average five-fold cross validation on the CASME II dataset for the best classifier i.e. Linear SVM with transfer learning was shown in Fig. 7. The diagonal elements were the proportion of facial micro-expressions that were correctly classified in the test set while the others were misclassified. The darker the color was, the higher the accuracy was. It can be seen that, our approach performed very well in recognizing disgust, repression, and surprise with recognition accuracy of 100%, while it was not perfect for happiness and others. The lowest accuracy was achieved by the emotion others with 89% and it was falsely classified to ‘disgust’ and ‘happiness’ categories with 9% and 2%, respectively. It may be explained that the ‘others’ was not an explicit facial micro-expression category and easily confused to similar categories.

Furthermore, the 3D-CNNs method was compared with state-of-art researches in recognizing facial micro-expression on CASME II (Table 4). Among these methods, CNN+LSTM [9], and TIM+DCNN+SVM [16] were deep learning based methods, while the others were traditional hand-craft methods, with diverse feature extraction manners.

Our 3D-CNNs algorithm achieved high accuracy of 97.6% with transfer learning in five-fold cross validation, and it was better than the performance of LBP-TOP-ELM [1], which applied the same cross-validation scheme on a small subset of CASME II. To further show the superiority of our approach, we also presented the results trained the model from scratch on CASME II dataset without transfer learning (with accuracy of 94.15%). The performance of our method with transfer learning was higher than performance of our method without transfer learning as 3.4% up. It can be found that the knowledge learned from source domain, which was the normal facial expression, made some contribution to the recognition of facial micro-expression.

Besides, the TIM+DCNN+SVM [16] utilized leave-one-sample-out validation while others used leave-one-
subject-out (LOSO) validation. Table 4 illustrated that the 3D-CNNs model achieved the highest accuracy of 65.9% with transfer learning, and 63.4% without transfer learning, in leave-one-subject-out cross validation. In HOG [11], the recognition accuracy for leave-one-sample-out was 57.5%, which was higher than the performance of leave-one-subject-out cross validation as 10% up. It was mainly because LOSO cross validation is stricter than leave-one-sample-out or other protocols. This phenomenon was consistent to the results in our experiments.

Moreover, CNN + LSTM [9] used data augmentation during the training to avoid overfitting, and 150 augmented sequences were generated from each training sequence. TIM+DCNN+SVM [16] adopted temporal interpolation model [22] to make the feature extraction simple. In our approach, the data augmentation was also conducted to avoid overfitting, and we obtained seven times more data combined with original sequences. Meanwhile, the 3D spline interpolation was adopted to normalize the input sequences, to make the spatiotemporal features more obvious. These methods also played important roles in improving the accuracy of facial micro-expression recognition.

4.4.2 SMIC

Experiments of five-fold cross validation and LOSO were conducted on the three subsets in SMIC, that was SMIC-HS, SMIC-VIS, and SMIC-NIR. Figure 8 illustrated the confusion matrix of the average five-fold cross validation on SMIC-HS for the best classifier of Linear SVM with transfer learning. The performance in the case of negative and positive was perfect (100%), while the performance for surprise was relatively poor (95%).

The comparison results on SMIC-HS were reported in Table 5. Among these methods, TIM+DCNN+SVM [16] was deep learning-based method, which obtained the best state-of-art accuracy of 65.8%. The recognition accuracy obtained in our LOSO validation experiment was 66.3% (with transfer learning) and 65.2% (without transfer learning), which was much higher than that reported in [16]. Moreover, in the person dependent experiments (five-fold cross validation), the average accuracy of 97.4% was achieved by 3D-CNNs plus transfer learning, which was more than twenty percent higher than person independent experiments (LOSO). In the results listed in Table 5, LBP-TOP+Nearest Neighbor [5] carried out person independent experiments and person dependent experiments, and the recognition accuracy of person dependent experiment was 65.8%, while the performance of person independent experiment was 53.7%.

Various experiments were also carried out on SMIC-VIS and SMIC-NIR sub-databases. And the performances of LOSO cross validation were compared with HOG [11], as it was the only literature in Table 5 that reported recognition results on these two subsets. On the SMIC-VIS, the 3D-CNNs model achieved recognition accuracy of 73.1% and 71.8% with or without transfer learning, respectively, while the results reported in [11] with LOSO was 71.8%. The
Table 5  Comparison of facial micro-expression recognition accuracy on SMIC-HS.

| Method                        | Accuracy |
|-------------------------------|----------|
| HOG [11]                      | 57.9%    |
| LBP-TOP+Nearest Neighbor [5]  | 65.8%    |
| LBP-TOP+GSLSR [13]            | 70.1%    |
| TIM+DCNN+SVM [16]             | 65.9%    |
| LOSO (train from scratch)     | 65.2%    |
| LOSO (with transfer learning) | 66.3%    |
| LBP-TOP+Nearest Neighbor [5]  | 53.7%    |
| Fivefold (train from scratch) | 95.8%    |
| Fivefold (with transfer learning) | 97.4%    |

result of 3D-CNNs model with transfer learning in SMIC-NIR database was 67.3% (LOSO) for three classes, which was 2% higher than the model without transfer learning, and almost 5% higher than that reported in [11].

As can be seen, the 3D-CNNs model with transfer learning scheme was superior to state-of-art approaches noticeably. This can be attributed to the interpolation of the facial micro-expression sequences and the efficient feature extraction ability of 3D-CNNs. The 3D-CNNs model was good at describing both texture and spatiotemporal features from the subtle facial micro-expression sequences. Transfer learning could well improve the recognition results of 3D-CNNs model due to better knowledge learned from source domain with sufficient samples, which dealt with the problem of limited labeled facial micro-expression samples in the available databases.

5. Conclusion

In this paper, we presented a novel deep learning based method for facial micro-expression recognition by combining 3D-CNNs with transfer learning. The facial micro-expression sequences were initially normalized by 3D spline interpolation and then the facial texture and motion characteristics were extracted using 3D-CNNs. To overcome the data limitation of facial micro-expression, the transfer learning from normal facial expression database was employed in our method. The proposed algorithm was evaluated on two available published facial micro-expression datasets, and our method had shown its superiority over traditional methods. Moreover, the 3D-CNNs had been proven to be well-performed for extracting the spatiotemporal features from the subtle facial micro-expression sequences. Besides, employing the transfer learning had leaded to a noticeable improvement in the accuracy of facial micro-expression recognition. Furthermore, we explored the effect of the pre-training in our experiments and the results demonstrated that the fine-tuned model performed best when we learned more generic features during the pre-training stage. In the future work, we will focus on facial micro-expression detection in videos, to further solve the problem of insufficient samples, and develop a comprehensive automatic facial micro-expression system, which would be helpful in practical applications.

Acknowledgments

This work is supported by the National Research and Development Major Project (2017YFD0400100), the National Natural Science Foundation of China (61673052), the Fundamental Research Fund for the Central Universities of China (2302018FRF-TP-18-014A2), and the grant from Chinese Scholarship Council (CSC).

References

[1] I.P. Adegun and H.B. Vadapalli, “Automatic recognition of micro-expressions using local binary patterns on three orthogonal planes and extreme learning machine,” Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech), pp.1–5, 2016.
[2] X. Duan, Q. Dai, X. Wang, Y. Wang, and Z. Hua, “Recognizing spontaneous micro-expression from eye region,” Neurocomputing, vol.217, pp.27–36, 2010.
[3] M.G. Frank, M. Herbasz, K. Sinuk, A. Keller, and C. Nolan, “I see
how you feel: Training laypeople and professionals to recognize communication emotions.” The Annual Meeting of the International Communication Association, Sheraton New York, New York City, 2009.

[4] D. Ghimire and J. Lee, “Geometric feature-based facial expression recognition in image sequences using multi-class adaboost and support vector machines,” Sensors, vol.13, no.6, pp.7714–7734, 2013.

[5] Y. Guo, Y. Tian, X. Gao, and X. Zhang, “Micro-expression recognition based on local binary patterns from three orthogonal planes and nearest neighbor method,” 2014 International Joint Conference on Neural Networks (IJCNN), pp.3473–3479, 2014.

[6] O.E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R.R. Salakhutdinov, “Improving neural networks by preventing co-adaptation of feature detectors,” Computer Science, vol.3, no.4, pp.212–223, 2012.

[7] X. Huang, G. Zhao, X. Hong, W. Zheng, and M. Pietikäinen, “Spontaneous facial micro-expression analysis using spatiotemporal completed local quantized patterns,” Neurocomputing, vol.175, pp.564–578, 2016.

[8] S. Ji, W. Xu, M. Yang, and K. Yu, “3D convolutional neural networks for human action recognition,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.35, no.1, pp.221–231, 2013.

[9] D.H. Kim, W.J. Baddar, and Y.M. Ro, “Micro-Expression Recognition with Expression-State Constrained Spatio-Temporal Feature Representations,” Proceedings of the 2016 ACM on Multimedia Conference, pp.382–386, 2016.

[10] D.P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” Computer Science, 2014.

[11] X. Li, X. Hong, A. Moilanen, X. Huang, T. Pfister, G. Zhao, and M. Pietikäinen, “Towards Reading Hidden Emotions: A Comparative Study of Spontaneous Micro-expression Spotting and Recognition Methods,” IEEE Transactions on Affective Computing, vol.9, no.4, pp.563–577, 2018.

[12] X. Li, T. Pfister, X. Huang, G. Zhao, and M. Pietikäinen, “A spontaneous micro-expression database: Inducement, collection and baseline,” 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), pp.1–6, 2013.

[13] P. Lu, W. Zheng, Z. Wang, Q. Li, Y. Zong, M. Xin, and L. Wu, “Micro-Expression Recognition by Regression Model and Group Sparse Spatio-Temporal Feature Learning,” IEICE Transactions on Information and Systems, vol.E99-D, no.6, pp.1694–1697, 2016.

[14] D. Matsumoto and H.S. Hwang, “Evidence for training the ability to read microexpressions of emotion,” Motivation and Emotion, vol.35, no.2, pp.181–191, 2011.

[15] D. Matsumoto and H.S. Hwang, “Reading facial expressions of emotion,” Psychological Science Agenda, vol.25, no.5, 2011.

[16] V. Mayya, R.M. Pai, and M.M.M. Pai, “Combining temporal interpolation and DCNN for faster recognition of micro-expressions in video sequences,” International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp.699–703, 2016.

[17] N. Michael, M. Diliszian, D. Metaxas, and J.K. Burgeon, “Motion profiles for deception detection using visual cues,” European Conference on Computer Vision, vol.6316, pp.462–475, 2010.

[18] V. Nair and G.E. Hinton, “Rectified linear units improve restricted boltzmann machines,” Proceedings of the 27th international conference on machine learning (ICML-10), pp.807–814, 2010.

[19] S.J. Pan and Q. Yang, “A survey on transfer learning,” IEEE Transactions on knowledge and data engineering, vol.22, no.10, pp.1345–1359, 2010.

[20] S. Park and D. Kim, “Subtle facial expression recognition using motion magnification,” Pattern Recognition Letters, vol.30, no.7, pp.708–716, 2009.

[21] D. Patel, X. Hong, and G. Zhao, “Selective deep features for micro-expression recognition,” 23rd International Conference on Pattern Recognition (ICPR), pp.2258–2263, 2016.

[22] T. Pfister, X. Li, G. Zhao, and M. Pietikäinen, “Recognising spontaneous facial micro-expressions,” IEEE International Conference on Computer Vision (ICCV), pp.1449–1456, 2011.

[23] M. Rasouzadeh, “Facial expression recognition using fuzzy inference system,” International Journal of Engineering and Innovative Technology, vol.1, no.4, pp.1–5, 2012.

[24] C. Shan, S. Gong, and P.W. McOwan, “Facial expression recognition based on local binary patterns: A comprehensive study,” Image and Vision Computing, vol.27, no.6, pp.803–816, 2009.

[25] X.-B. Shen, Q. Wu, and X.-L. Fu, “Effects of the duration of expressions on the recognition of micro-expressions,” Journal of Zhejiang University-Science B, vol.13, no.3, pp.221–230, 2012.

[26] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, “Learning spatiotemporal features with 3d convolutional networks,” Proceedings of the IEEE international conference on computer vision, pp.4489–4497, 2015.

J. Tunney, “Tensorflow,” available: https://github.com/tensorflow/tensorflow/tree/r0.10

S.-J. Wang, W.-J. Yan, T. Sun, G. Zhao, and X. Fu, “Sparse tensor canonical correlation analysis for micro-expression recognition,” Neurocomputing, vol.214, pp.218–232, 2016.

W.-J. Yan, X. Li, S.-J. Wang, G. Zhao, Y.-J. Liu, Y.-H. Chen, and X. Fu, “CASME II: An improved spontaneous micro-expression database and the baseline evaluation,” PLoS One, vol.9, no.1, pp.e86041, 2014.

W.-J. Yan, S.-J. Wang, Y.-J. Liu, Q. Wu, and X. Fu, “For micro-expression recognition: Database and suggestions,” Neurocomputing, vol.136, pp.82–87, 2014.

W.-J. Yan, Q. Wu, J. Liang, Y.-H. Chen, and X. Fu, “How fast are the leaked facial expressions: The duration of micro-expressions,” Journal of Nonverbal Behavior, vol.37, no.4, pp.217–230, 2013.

Z. Zeng, M. Pantic, G.I. Roisman, and T.S. Huang, “A survey of affect recognition methods: Audio, visual, and spontaneous expressions,” IEEE transactions on pattern analysis and machine intelligence, vol.31, no.1, pp.39–58, 2009.

G. Zhao, X. Huang, M. Taini, S.Z. Li, and M. Pietikäinen, “Facial expression recognition from near-infrared videos,” Image and Vision Computing, vol.29, no.9, pp.607–619, 2011.

R. Zhu, T. Zhang, Q. Zhao, and Z. Wu, “A transfer learning approach to cross-database facial expression recognition,” International Conference on Biometrics (ICB), pp.293–298, 2015.

X. Huang, S.-J. Wang, G. Zhao, and M. Pietikäinen, “Facial micro-expression recognition using spatiotemporal local binary pattern with integral projection,” IEEE International Conference on Computer Vision, Santiago, Chile, 2015.

Y. Wang, J. See, Y.-H. Oh, R.C.-W. Phan, Y. Rahulamathavan, H.-C. Ling, S.-W. Tan, and X. Li, “Effective recognition of facial micro-expression with video motion magnification,” Multimedia Tools and Applications, vol.76, no.20, pp.21665–21690, 2017.

Ruicong Zhi received the PhD degree in signal and information processing from Beijing Jiaotong University in 2010. From 2016–2017, she visited the University of South Florida as a visiting scholar. She visited the Royal Institute of Technology (KTH) in 2008 as a joint PhD. She is currently an associate professor in School of Computer and Communication Engineering, University of Science and Technology Beijing. She has published more than 50 papers, and has six patents. She has been the recipient of more than ten awards, including the National Excellent Doctoral Dissertation Award nomination. Her research interests include facial and behavior analysis, artificial intelligence, and pattern recognition. 
Hairui Xu received the MS degree at School of Computer and Communication Engineering from University of Science and Technology Beijing in 2018, and received the Bachelor degree in Computer Science from University of Science and Technology Beijing in 2015. His research interest includes computer vision and emotion analysis.

Ming Wan received the Bachelor degree at School of Information Science and Technology from University of Beijing Forestry University in 2017. She is currently pursuing the MS degree at School of Computer and Communication Engineering, University of Science and Technology Beijing. Her research interest includes computer vision and deep learning.

Tingting Li received the Bachelor degree in Computer Science from University of Science and Technology Beijing in 2016. She is currently pursuing the MS degree at School of Computer and Communication Engineering, University of Science and Technology Beijing. Her research interest includes computer vision and emotion analysis.