Article

Facial Micro-Expression Recognition based on Deep Local-Holistic Network

Jingting Li [1], Ting Wang [2] and Su-Jing Wang [1,3,*]

[1] CAS Key Laboratory of Behavioral Science, Institute of Psychology
[2] Department of Computer and Information Technology, Beijing Jiaotong University
[3] Department of Psychology, University of the Chinese Academy of Sciences
* Correspondence: wangsujing@psych.ac.cn.

Abstract: Micro-expression is a subtle, local and brief facial movement. It can reveal the genuine emotions that a person tries to conceal and is considered an important clue for lie detection. The micro-expression research has attracted much attention due to its promising applications in various fields. However, due to the short duration and low intensity of micro-expression movements, micro-expression recognition faces great challenges, and the accuracy still demands improvement. To improve the efficiency of micro-expression feature extraction, inspired by the psychological study of attentional resource allocation for micro-expression cognition, we propose a deep local-holistic network method for micro-expression recognition. Our proposed algorithm consists of two sub-networks. The first is a Hierarchical Convolutional Recurrent Neural Network (HCRNN), which extracts the local and abundant spatio-temporal micro-expression features. The second is a Robust principal component analysis-based recurrent neural network (RPRNN), which extracts global and sparse features with micro-expression-specific representations. The extracted effective features are employed for micro-expression recognition through the fusion of sub-networks. We evaluate the proposed method on combined databases consist of four most commonly used databases, i.e., CASME, CASME II, CAS(ME)^2, and SAMM. The experimental results show that our method achieves a reasonably good performance.

Keywords: Hierarchical Convolution; Local-Holistic; micro-expression recognition; Robust Principal Component Analysis

Citation: Li, J.; Wang, T.; Wang, S-J. Facial Micro-Expression Recognition based on Deep Local-Holistic Network. *Appl. Sci.* **2022**, *1*, 0. https://doi.org/

Received:
Accepted:
Published:

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

Facial micro-expression (micro-expression) is an involuntary and momentary facial expression, with a brief duration of less than 500ms [1]. It reflects one's genuine emotions that people are trying to conceal. In contrast to ordinary facial expressions, micro-expression is consciously suppressed, but unconsciously leaked. Moreover, it has the two distinguishing features of short duration and low intensity. Compared to polygraph instruments that require equipment, micro-expression-based lie detection is unobtrusive, and individuals are less likely to counteract it. Therefore, micro-expressions have many potential applications in many fields, such as clinical diagnosis [2] and national security [3].

micro-expression is difficult to detect through the naked eye and requires a trained professional to recognize [2]. In order to help people recognize micro-expression, Ekman *et al.* developed the Facial Action Coding System (FACS) [4] and defined the muscle activity of facial expressions as action units (AU). Meantime, they also developed the micro-expression Training Tool (micro-expressionTT) [5]. Since then, micro-expression has received increasing attention from researchers. However, micro-expression analysis through humans is still very challenging, and many researchers have tried to develop micro-expression auto-recognition methods by employing computer vision techniques. Since 2013, Xiaolan Fu's group has built three spontaneous micro-expression databases:

CASME I [6], CASME II [7], CAS(ME)^2 [8]. In 2016, Davison *et al.* released the Spontaneous Actions and Micro-Movements (SAMM) [9] dataset with demographical diversity.

Based on these published databases, research on micro-expression recognition has been gradually developed. There are two main types of approaches, i.e., recognition methods based on handcraft features and methods based on deep learning feature extraction. Due to the brief, subtle, and localized nature of micro-expressions, it is challenging for both handcrafted features and features obtained based on deep learning to fully represent micro-expressions. In addition, since the collection and labeling of micro-expressions are time-consuming and laborious, the total number of published micro-expression samples is about 1000. Therefore, micro-expression recognition is a typical small sample size (SSS) problem. The sample size greatly limits the application of deep learning in this area. First, deep network models involve a large number of parameters, and training on a small micro-expression sample may cause overfitting problems of the model. Moreover, the number of samples in the model and the network parameters are affected by the SSS problem compared with the algorithms for expression recognition. Furthermore, due to the complicated characterization of micro-expressions themselves, even methods such as transfer learning with sample pre-training on other large-scale data sets do not achieve satisfactory results cannot be applied to practical applications.

To address the problem that micro-expression features are difficult to learn in deep networks under small sample problems, we explored the psychological cognitive attention mechanism. As shown in Fig. 1, the process of individual cognitive micro-expressions moves from global cognition to local-focused attention and finally to global decision making [10]. Inspired by this theory, we propose a Deep Local-Holistic Network (DLHN) with enhanced micro-expression feature extraction capability for micro-expression recognition. The architecture of the proposed method mainly includes two sub-networks: (1) a hierarchical convolutional recurrent network (HCRNN), learning local and abundant features from original frames of micro-expression video clips. (2) a robust principal component analysis recurrent network (RPRNN), extracting sparse information from original frames of micro-expression video clips by RPCA, and then feeding the sparse information to a deep learning model to extract holistic and sparse features. The two networks are trained separately and then fused for micro-expression recognition.

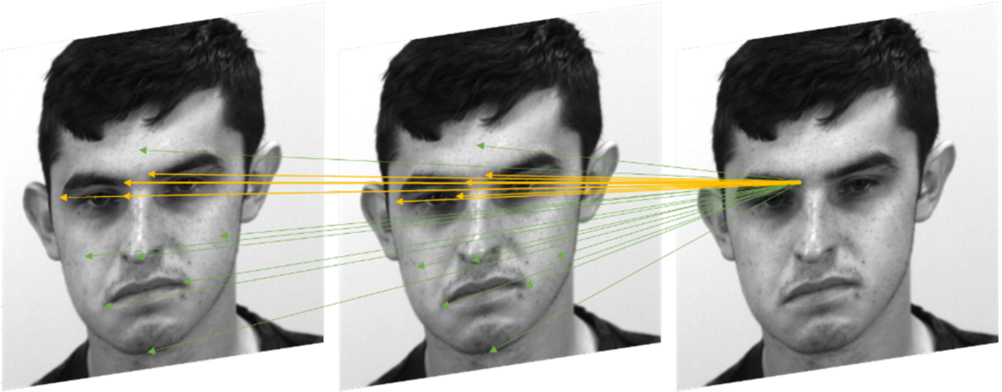


Figure 1. Global (green clipping head) and local area of interest (yellow arrow) tracking of micro-expression action. (Sample from SAMM dataset)

The rest of this paper is organized as follows: Section 2 reviews the related works on micro-expression recognition and basic models applied in our method; Section 3 introduces our proposed algorithm in detail; Section 4 presents the experimental results; and Section 5 concludes the article.

2. Related Works and Background

This section first introduces the related works on micro-expression recognition, then briefly describes three algorithms as they are employed in our proposed method, including deep convolutional neural network, recurrent neural network, and Robust Principal Component Analysis.

2.1. micro-expression Recognition

In the early stages of the study, most methods adopt handcrafted features to identify micro-expressions. Polikovsky *et al.* [11] divided the face into specific regions and recognized the motion in each region by 3D-Gradients orientation histogram descriptor. Tomas Pfister *et al.* [12] designed the first spontaneous micro-expression database (SMIC) and used LBP-TOP [13] to extract dynamic and appearance features of micro-expressions. Wang *et al.* [14] adopted robust Principal Component Analysis (RPCA) [15] to extracted sparse micro-expression information and Local Spatiotemporal Directional Features. Wang *et al.* introduced a discriminant tensor subspace analysis (DTSA) [16] to preserve the spatial structure information of micro-expression images. Furthermore, they treated micro-expression video clip as a fourth-order tensor and transformed the color information from RGB into TICS to improve the performance [17]. Huang *et al.* [18] show a spatiotemporal facial representation to characterize facial movements and used LBP to extract appearance and motion features. Liu *et al.* [19] proposed a simple, effective Main Directional Mean Optical-flow features (MDMO) and adopted SVM classifier to recognize micro-expression. Huang *et al.* [20] analyzed micro-expression by proposing SpatioTemporal Completed Local Quantization Patterns (STCLQP), which exploits magnitude and orientation as complementary features. The above recognition methods are not capable enough to capture subtle facial displacements. This is due to the constant movement of the observed individual, which is common in typical micro-expression applications. Addressing this problem, Xu *et al.* [21] proposed a Facial Dynamics Map method with depicting micro-expression characteristics from different granularity. Wang *et al.* [22] proposed a Main Directional Maximal Difference micro-expression recognition method (MDMD), extracting optical flow features from the region of interest (ROIs) based on action units.

Recently, the outstanding performance of deep learning attracts the attention of many researchers to develop micro-expression recognition algorithms. Patel *et al.* [23] used the pre-trained ImageNet-VGG-f CNN to extract features of each frame in micro-expression video clips. Wang *et al.* [24] proposed a Transferring Long-term Convolutional Neural Network (TLCNN) method, which uses Deep CNN to extract spatial features per frame and Long Short Term Memory (LSTM) to learn micro-expression temporal information. Xia *et al.* [25] investigated a low-complexity recurrent convolutional neural network (RCN) based on cross-database micro-expression recognition. Li *et al.* [26] performed a joint local and global information learning on apex frame for micro-expression recognition. Zhou *et al.* [27] proposed an expression-specific feature learning and fusion method for micro-expression recognition However, the small sample size of micro-expression samples and the subtle and brief nature of micro-expression limit the combination of deep learning with micro-expression recognition methods. Thus, how to learn the micro-expression features effectively is necessary research for further performance improvement.

2.2. Deep Convolutional Neural Network

Deep Convolutional neural network (DCNN) is a hierarchical machine learning method containing multilevel nonlinear transformations. It is a classic and widely used structure with three prominent characteristics: local receptive fields shared weights and spatial or temporal subsampling. These features reduce temporal and spatial complexity and allow some degree of shift, scale, and distortion invariance when designed to process still images. It has been shown to outperform many other techniques [28].

As introduced in Section 1, the handcraft micro-expression features are not sufficiently representational. Hence, we apply DCNN to improve the discriminative ability for micro-expression by targeting learning in local regions where micro-expressions frequently occur.

2.3. Recurrent Neural Network

Recurrent neural network (RNN) can be used to process sequential data through mapping an input sequence to a corresponding output sequence, using the hidden states. However, as the network gradually deepens, there will be problems of gradient disappearance and gradient explosion. To solve this problem, Long Short-Term Memory (LSTM) architecture was proposed [29] which uses memory cells with multiplicative gate units to process information. It has been shown to outperform RNN on learning long sequences.

Besides, RNN takes into account only the past context. To solve the problems, a bidirectional RNN (BRNN) is created [30], which can process data in both past and future information. Subsequently, Graves *et al.* [31] proposed a bidirectional LSTM (BLSTM), which has better performance than LSTM on processing long contextual information of complex temporal dynamics.

Since micro-expressions are very subtle, it isn't easy to distinguish them from neutral faces just by a single frame. The movement pattern in the temporal sequence is an essential feature for micro-expressions. Therefore, we extract the temporal features from micro-expression sequence based on BRNN and BLSTM to enhance the classification performance.

2.4. Robust Principal Component Analysis

Donoho *et al.* [32] demonstrated that the observed data could be separated efficiently and exactly into sparse and low-rank structures in high-dimensional spaces. Then, an idealized "robust principal component analysis" problem is proposed to recover a low-rank matrix A from highly corrupted measurements D:

$$\mathbf{D} = \mathbf{A} + \mathbf{E} \tag{1}$$

Where **A** is the deserved data in a low-rank subspace, and **E** is the error term, usually treated as noise.

According to the characteristic of micro-expression with short duration and low intensity, micro-expression data are sparse in both the spatial and temporal domains. In 2014, Wang et.al. [17] proposed **E** as the deserved subtle motion information of micro-expression and **A** as noise for micro-expression recognition. Inspired by this idea, we adopt RPCA to obtain sparse information from micro-expression frames, and then feed the extracted information to RPRNN which learns sparse and holistic micro-expression features.

3. Our Model

As illustrated in Fig. 2, our proposed Deep Local-Holistic Network (DLHN) consists of HCRNN and RPRNN. HCRNN extracts the local and abundant spatial-temporal micro-expression features by concatenating modified CNN and BRNN modules. Meanwhile, RPRNN learns the holistic sparse micro-expression features through the combination of RPCA and a deep BLSTM. Finally, two sub-networks are fused to improve the performance of micro-expression recognition.

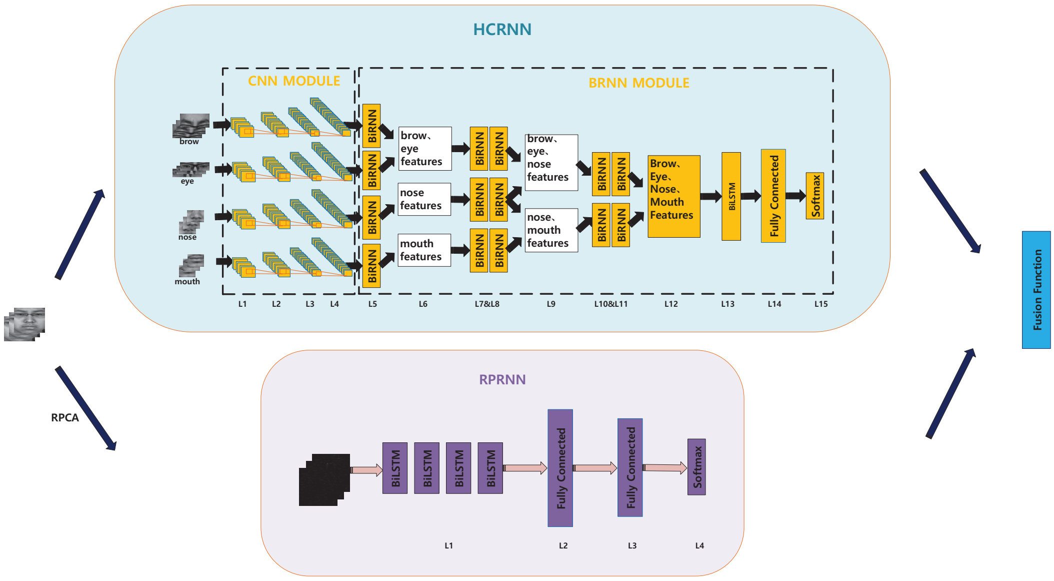


Figure 2. Our proposed Deep Local-Holistic Network. (1) The local network, i.e., HCRNN. The facial micro-expression image is divided into four regions of interest and then fed into four hierarchical CNN modules to extract local-still features. In addition, local dynamic features are learned by the BRNN module. (2) The holistic network, i.e., RPRNN. RPCA is employed to obtain sparse micro-expression images, which are then used as the input to the RPRNN. A deep BLSTM network created by multiple hidden layers is applied to learn the holistically sparse features.

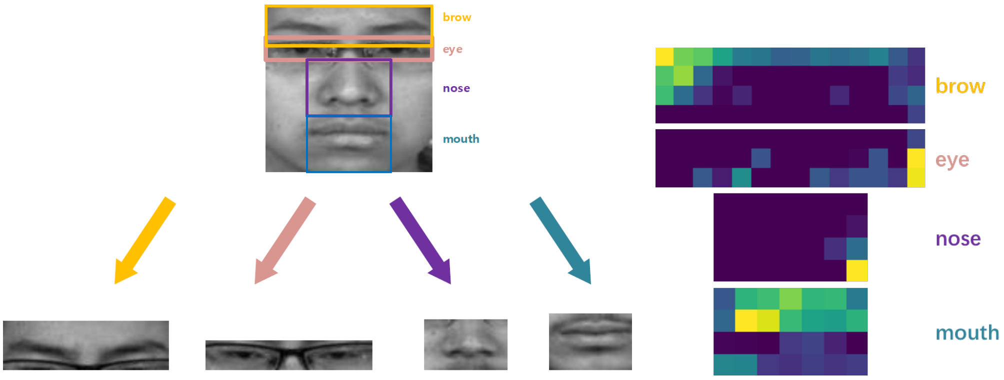
3.1. HCRNN for Local Features

As illustrated in the top block of Fig. 2, the HCRNN Model is constructed by CNN Module and BRNN Module. First, CNN Module contains four hierarchical CNNs (HCNNs) to extract local features from ROIs. Then, BRNN Module learns the temporal correlation in the local features. Finally, the category of micro-expression is predicted by a fully connected (FC) layer.

3.1.1. CNN Module

According to the facial physical structure, only four facial regions of interest (ROIs), i.e., eyebrows, eyes, nose, mouth, are used for the local micro-expression feature extraction (Fig. 3a). First, the gray-scale micro-expression frames are cropped and normalized with a size of 112×112. Then the ROIs are determined based on facial landmarks. The ROI size of eyebrows, eyes, nose and mouth are 112 ×33, 112×20, 56×32, 56×38, respectively. Furthermore, considering the integrity of each ROI, the adjacent ROIs may have overlapping portions.

As shown in the HCRNN bock of Fig. 2, the structure of CNN module consists of four HCNNs. For each branch, the input is the ROI gray-scale images, and the network contains four convolutional layers. All four HCNNs have the same structure, as listed in Table 1. The output sizes in the table refer to generated tensor shapes by four HCNN. The CNN module is able to extract local spatial micro-expression features. For a better visualization, Fig. 3b presents the feature maps of L4 in HCRNN.



(a) Four facial ROIs **(b)** ROI feature maps

Figure 3. ROIs based on eyebrows, eyes, nose and mouth, and the corresponding feature maps of L4 in HCRNN.

Table 1. The HCNN structure

Type	Kernel size	Stride	Output size
convolution	3 × 3 × 70	1	112 × 33/112 × 20/56 × 32/56 × 38
max pool	2 × 2	2	56 × 16/56 × 10/28 × 16/28 × 19
convolution	3 × 3 × 140	1	56 × 16/56 × 10/28 × 16/28 × 19
max pool	2 × 2	2	28 × 8/28 × 5/14 × 8/14 × 9
convolution	3 × 3 × 280	1	28 × 8/28 × 5/14 × 8/14 × 9
max pool	2 × 2	2	14 × 4/14 × 2/7 × 4/7 × 4
convolution	3 × 3 × 560	1	14 × 4/14 × 2/7 × 4/7 × 4

3.1.2. BRNN Module

In a micro-expression sequence, the past context and future context usually are useful for prediction. Thus, a BRNN module [33] is adopted to process micro-expression temporal variation. The number of neurons in each layer of BRNN Module is listed as follows: L5(30 × 4)-L7(60 × 3)-L8(60 × 3)-L10(90 × 2)-L11(90 × 2)-L12(80 × 1). First, the extracted ROI features from CNN module are fed into BRNN module in L5 layer. Then, local temporal information is concatenated in L6 layer and subsequently processed by two BLSTMs in L7 layers (See BRNN structure in Fig. 4). Similar steps of L6 and L7 are repeated in L8 and L9 layers. A global temporal feature is obtained through the concatenation in L10 layer and the BLSTM in L11 layer. We classify micro-expression by an FC layer in L12 of HCRNN and obtain probabilistic outputs by softmax layer in L13 of HCRNN:

$$\mathrm{P}(\mathrm{h_i}) = \frac{e^{h_i}}{\sum_{k=0}^{n-1} e^{h_k}} \tag{2}$$

where h_i is the output of L13, i is the output unit, where $i = 0, 1, ...k$. Finally, the HCRNN is trained by using the cross-entropy loss function:

$$\mathrm{HLoss} = -\sum_{j} \mathrm{c_j} \cdot \log(\mathrm{P}(\mathrm{h_j})) \tag{3}$$

where c_j is the ground truth, $\mathrm{P}(\mathrm{h_j})$ is the predicted probability of output layer.

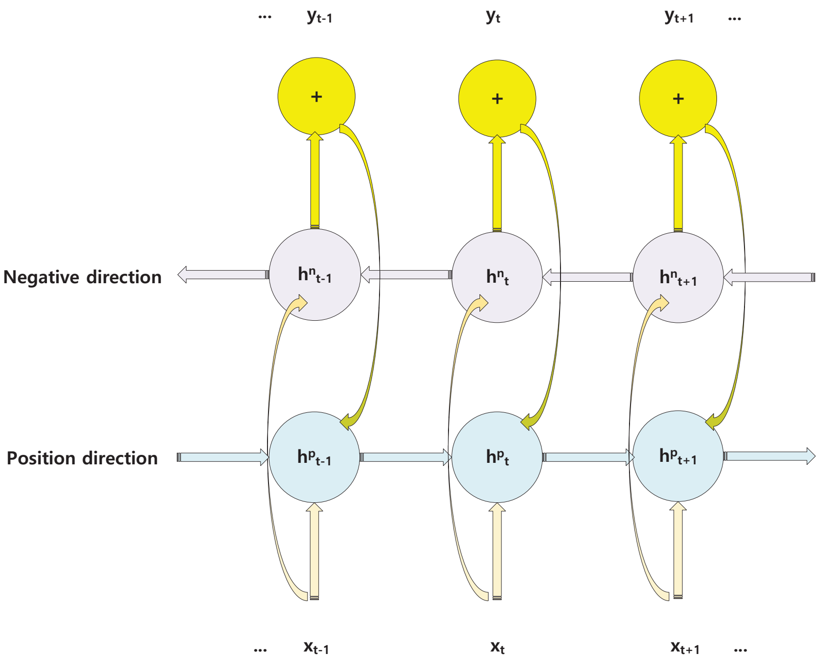

... y_{t-1} y_t y_{t+1} ...
+
Negative direction
h^n_{t-1} h^n_t h^n_{t+1}
Position direction
h^p_{t-1} h^p_t h^p_{t+1}
... x_{t-1} x_t x_{t+1} ...

Figure 4. General structure of BRNN. x_t is input data in t time. y_t is output data in t time. h_t^p and h_t^n represent the hidden state in positive and negative directions, respectively.

3.2. RPRNN for Holistic Features 181

3.2.1. Input: Sparse micro-expression Obtained by RPCA 182

Due to the short duration and low intensity of micro-expression movement, micro-expression could be considered as sparse data. Hence, RPCA [15] is utilized to obtain sparse micro-expression information. In details, for a gray-scale video clip $\mathcal{V}(h \times w \times n)$, where h and w is respectively the pixels height and width of each frame, n is the number of frames. We stack all frames as column vectors of a matrix D with $h \times w$ rows and n columns. It can be formulated as follows:

$$\min_{\mathbf{A},\mathbf{E}} \operatorname{rank}(\mathbf{A}) + \|\mathbf{E}\|_0 \quad \text{subject to} \quad \mathbf{D} = \mathbf{A} + \mathbf{E} \tag{4}$$

where A is a low-rank matrix, B is a sparse matrix, $\operatorname{rank}(\cdot)$ is the rank of the matrix and $\|\cdot\|_0$ denotes ℓ_0 -norm which obtains the number of nonzero elements in the matrix. This is a non-convex function. Wright *et al.* adopted the ℓ^1 -norm as a convex surrogate for the highly-nonconvex ℓ^0 -norm and the nuclear norm (or sum of singular values) to replace non-convex low-rank matrix, i.e., the following convex optimization problem:

$$\min_{\mathbf{A},\mathbf{E}} \|\mathbf{A}\|_* + \lambda\|\mathbf{E}\|_1 \quad \text{subject to} \quad \mathbf{D} = \mathbf{A} + \mathbf{E} \tag{5}$$

where $\|\cdot\|_*$ denotes nuclear norm, $\|\cdot\|_1$ denotes ℓ_1 -norm which counts the sum of all elements in matrix, and λ is a positive weighting parameter ($\lambda > 0$). Lin *et al.*[34] proposed the Augmented Lagrange Multiplier Method (ALM), which includes two algorithms of exact ALM and inexact ALM to process linearly constrained convex optimization problems. The inexact ALM has a slight improvement in the required number of partial SVDs than the exact ALM and has the same convergence speed as the exact ALM. Benefiting from it, we adopt the method of inexact ALM to obtain sparse micro-expression motion information from original frames.

3.2.2. RPRNN Architecture 191

The obtained sparse micro-expression images are fed into RPRNN to extract holistic features. The architecture of RPRNN is shown at the bottom block in Fig. 2. In order to learn high-level micro-expression representations, a deep BLSTM network is created by multiple LSTM hidden layers. The holistically sparse features are extracted in the L1 of RPRNN, and two FC layers are used to classify micro-expressions. Then, the emotion type of micro-expression is estimated by the softmax layer:

$$\mathrm{P}(\mathrm{r_i}) = \frac{e^{r_i}}{\sum_{k=0}^{C-1} e^{r_k}} \tag{6}$$

where r_i is an output of the softmax layer. Finally, to avoid the overfitting problem, we combine the cross-entropy loss function with L2 Regularization:

$$\mathrm{RLoss} = -\sum_{j} \mathrm{c_j} \cdot \log\left(\mathrm{P}(\mathrm{r_j})\right) + \sum_{c=1}^{n} \theta_c^2 \tag{7}$$

where $\mathrm{P}(\mathrm{r_i})$ is the predicted probability of output layer, θ index to weight values. 192

3.3. Model Fusion 193

In the final stage of our proposed Deep Local-Holistic Network, HCRNN and RPRNN are fused by the following function:

$$O(x_i) = aP_{hi}(x_i) + (1-a)P_{ri}(x_i) \tag{8}$$

where a is weight value, P_{hi} and P_{ri} are the predicted probabilities in HCRNN and RPRNN. 194
According to the experiment result, we find that the model can achieve the best performance 195
when a equals 0.45. Thus, we set a to 0.45. 196

4. Experiments 197

4.1. Databases and Protocols 198

We use the datasets combined of four spontaneous micro-expression databases (CASME I, CASME II, CAS(ME)^2, and SAMM) to assess the performance of our models. Table 2 presents the details of these four databases. However, the number of emotion classes number is different in these databases, and micro-expression samples are labeled by taking different AUs criteria. For example, the combination of AU1 and AU2 defines a micro-expression sample as disgust in CAS(ME)^2 and as surprise in CASME II. In order to alleviate the impact of the different encoding, we adopt a uniformly AU encoding criterion proposed by Davison *et al.* [35]. Finally, we select 560 samples from the combined dataset and divide them into four emotion labels:

$$\text{emotions} = \{\text{Positive}, \text{Negative}, \text{Surprise}, \text{Others}\} \tag{9}$$

Specifically, Negative Consists of anger, disgust, sadness, and fear. Fig.7a shows the sample 199
size of each emotion category. In our experiments, we use 10-fold cross-validation protocol 200
on our combined dataset. 201

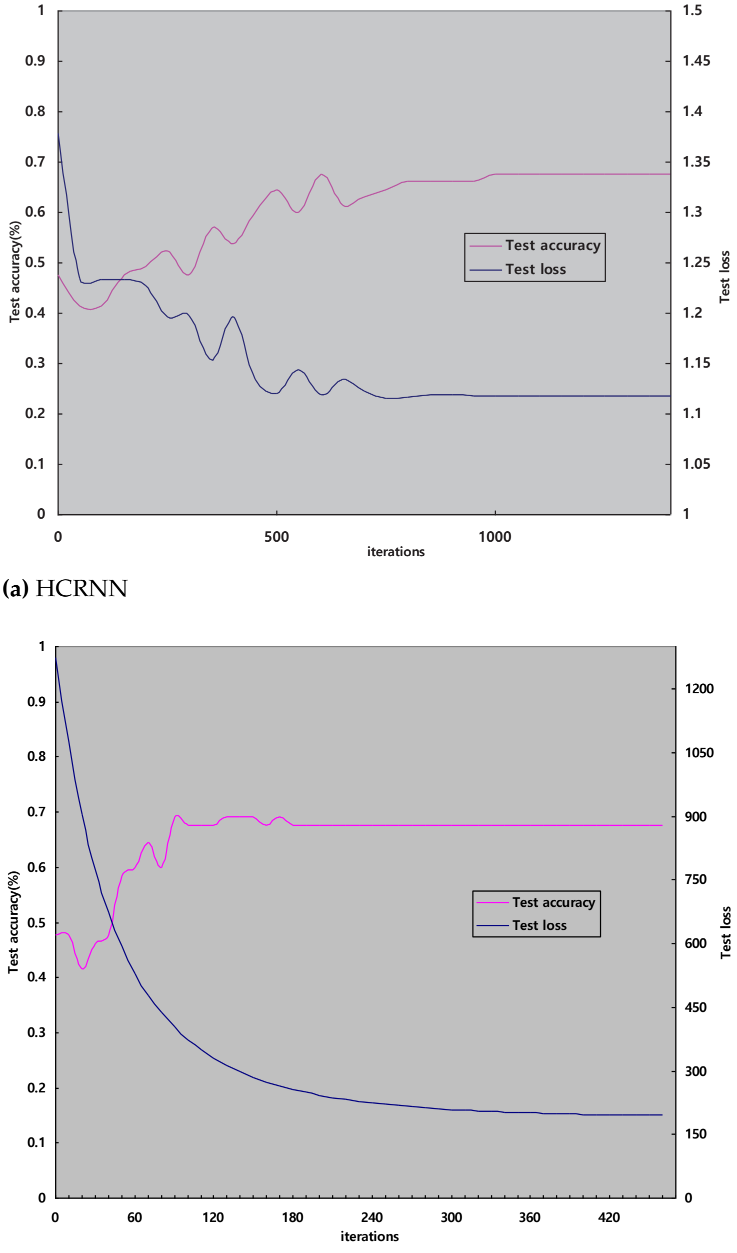
Table 2. Four spontaneous micro-expression databases. FPS: Frames per second

Database	Sample size	Emotions class	FPS	label
CASME I	195	8	60	emotion/AUs
CASME II	247	5	200	emotion/AUs
CAS(ME)^2	57	4	30	emotion/AUs
SAMM	159	7	200	emotion/AUs

4.2. Preprocessing and Parameter Configuration

Since the length of each video sample varies, we performed linear interpolation and extracted 16 frames from it for the subsequent recognition task. The size of the face image is 112 × 112. For HCRNN, the face region is divided into four ROIs as the input of CNN module. To guarantee the integrity of each part, ROIs have overlapping areas, and the size of brow, eye, nose, mouth regions are 112 × 33, 112 × 20, 56 × 32, and 56 × 38, respectively. The convolution kernel size of HCNN is set to 3 × 3, and the size of the pooling kernel is 2 × 2. The stride of convolution and pooling layer is set as 1 and 2. In the training stage, the learning rate adopts exponential decay with that the initial value equals 0.85 . We update all weights in each iteration with mini-batch samples whose size is 45. The iteration curves in Fig. 5a respectively represent the trend of loss and accuracy value in the testing set.

For RPRNN, the original micro-expression frames are processed by RPCA to obtain the sparse micro-expression images. Fig. 6 illustrates an example of micro-expression images processed by RPCA. Then the sparse images are fed to RPRNN to obtain holistic features. In the model, the attenuation way of learning rate and the update mode of weights are the same as HCRNN, and the value of the learning rate is initialized to 0.01. Same as HCRNN, in the training stage, we update all weights in each iteration. Fig.5b plots the iteration curves representing the trend of loss and accuracy value in the testing set. In the whole experiment, we employ a truncated normal distribution with zero mean and a standard deviation of 0.1 to initialize weights, and initialize biases as 0.1.



(a) HCRNN

(b) RPRNN

Figure 5. Network iteration curves

4.3. Results 222

Our proposed DLHN consists of HCRNN and RPRNN. As introduced in Section 3.3, these two sub-networks are combined by parameter a. We choose different a to evaluate the results of the fusion network and conduct our experiments with 10-fold cross-validation. Table 3 show micro-expression recognition accuracy of the fusion network with different parameter a. It can be seen that when a equals 0.45, the average accuracy of the fusion network is the highest. Therefore a is set as 0.45 when we compare the performance of the proposed DLHN with current state-of-the-art (SOTA) methods in the combined dataset.

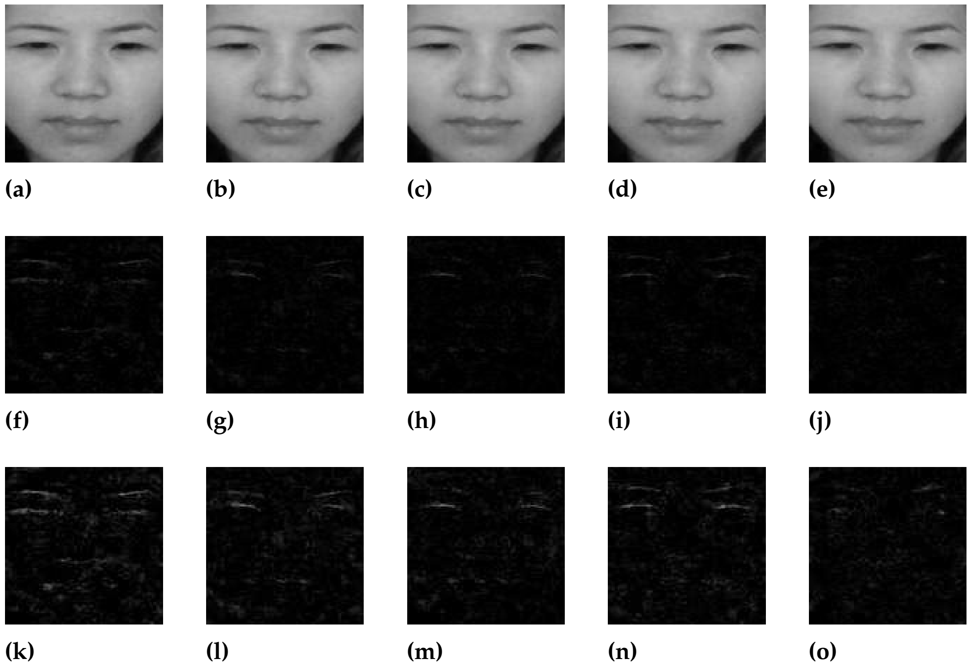


Figure 6. An example of RPCA on micro-expression images. Fig. 6a-6e are the original micro-expression images. Fig. 6f-6j are the corresponding extracted sparse information. Fig. 6k-6o are the enhanced display for Fig. 6f-6j by multiplying each pixels with 2.

Table 3. Facial micro-expression recognition accuracy (%) of our proposed DLHN with different parameter a in 10-fold cross-validation dataset

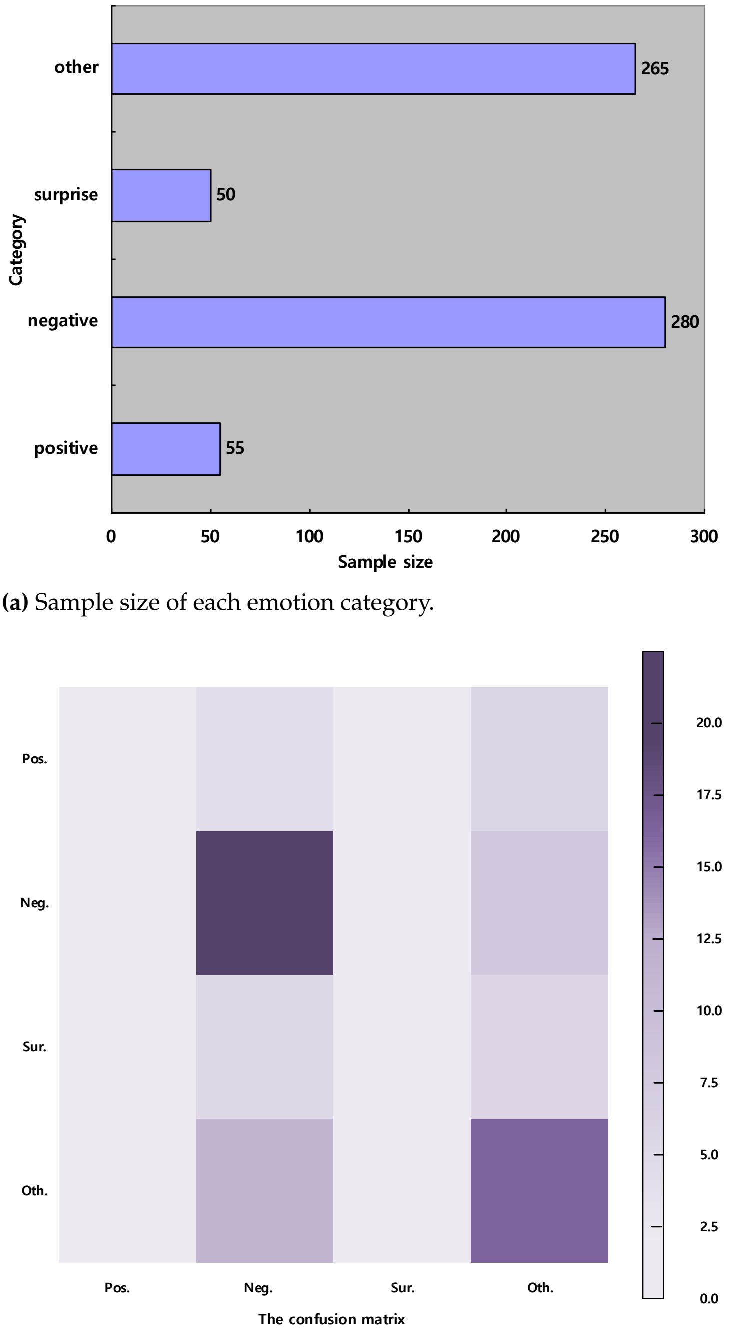
a	0.1	0.2	0.3	0.4	0.45	0.5	0.6	0.7	0.8	0.9
Fold1	55.38	52.31	55.38	53.85	53.85	**58.46**	**58.46**	56.92	55.38	53.85
Fold2	66.15	64.62	69.23	**70.77**	**70.77**	**70.77**	69.23	67.69	63.08	63.08
Fold3	60	60	61.54	61.54	61.54	**63.08**	61.54	63.08	60	58.46
Fold4	61.54	63.08	63.08	**66.15**	**66.15**	64.62	**66.15**	66.15	64.62	63.08
Fold5	56.92	56.92	55.38	**60**	**60**	58.46	58.46	58.46	56.92	56.92
Fold6	63.08	63.08	**64.62**	63.08	**64.62**	63.08	58.46	61.54	61.54	60
Fold7	**55.38**	53.85	52.31	53.85	47.69	41.54	41.54	41.54	41.54	41.54
Fold8	**60**	58.46	**60**	**60**	58.46	58.46	53.85	52.31	50.77	52.31
Fold9	52.31	52.31	52.31	53.85	53.85	**56.92**	53.85	53.85	**56.92**	**56.92**
Fold10	**63.08**	**63.08**	**63.08**	61.54	60	61.54	52.31	52.31	52.31	52.31
Mean	59.385	58.769	59.692	60.308	**60.309**	60.308	57.385	57.385	56.308	55.847

In the choice of comparison methods, among the handcraft feature-based methods, we choose the classical FDM features and LBP features [36], as well as the variant of LBP features (LBP-SIP) [37]. Among the deep learning methods, we choose the first place method for Micro-Expression Grand Challenge 2019 and two deep learning-based methods with codes released in the last two years, which are STSTNet [38], RCN(_a,_w,_s, and _f) [25] and Feature Refinement (FR) [27], respectively. Moreover, we all reproduced these methods with the same data configuration. Table 4 shows the overall accuracy of all algorithms. The best algorithm based on traditional methods for micro-expression recognition is LBP-TOP(4×4), which achieves 58.38% mean accuracy. The mean accuracy of HCRNN and RPRNN is respectively 55.08% and 59.53%. The fusion model, i.e., DLHN obtains the best performance by combined local abundant features extracted by HCRNN and holistic sparse features extracted by RPRNN and achieves 60.31% mean accuracy. Besides, RPRNN obtain the best performances in three folds (fold7, fold8, and fold10), which demonstrate that the efficiency of holistic sparse spatio-temporal feature extraction capacity of RPRNN.

Table 4. The overall accuracy (%) of DLHN and other SOTA methods. LBP_1, LBP_2, LBP_3 and LBP_4 reprensent LBP-TOP(2×2), LBP-TOP(4×4), LBP-SIP(2×2) and LBP-SIP(4×4) respectively.

Method	Fold1	Fold2	Fold3	Fold4	Fold5	Fold6	Fold7	Fold8	Fold9	Fold10	Mean
FDM+SVM	36.92	41.54	52.31	43.08	33.85	43.08	43.08	52.31	33.85	50.77	43.08
LBP_1+SVM	55.38	52.31	50.77	56.92	58.46	47.69	53.85	55.38	56.92	53.85	53.85
LBP_2+SVM	**66.15**	58.46	64.62	58.46	**63.08**	58.46	49.23	52.31	**61.54**	**61.54**	58.38
LBP_3+SVM	55.38	58.46	58.46	53.85	46.15	50.77	43.08	58.46	58.46	53.85	43.08
LBP_4+SVM	60	55.38	41.54	49.23	60	47.69	46.15	55.38	55.38	49.23	46.15
STSTNet	46.15	60.00	58.46	55.38	50.77	53.85	50.77	49.23	52.31	55.38	53.23
RCN_w	47.69	61.54	53.85	52.31	49.23	56.92	46.15	58.46	55.38	53.85	53.54
RCN_s	38.46	63.08	49.23	56.92	46.15	53.85	46.15	55.38	60.00	36.92	50.61
RCN_a	35.38	61.54	47.69	61.54	46.15	46.15	49.23	64.62	47.69	36.92	49.69
RCN_f	46.15	72.31	56.92	53.85	46.15	50.77	53.85	50.77	58.46	47.69	53.69
FR	46.15	61.54	58.46	**66.15**	61.54	56.92	50.77	44.62	56.92	56.92	56.00
HCRNN	53.85	63.08	58.46	63.08	56.92	56.92	40	52.31	55.38	50.77	55.08
RPRNN	56.82	64.62	60	61.54	56.92	60	**56.92**	**60**	56.92	**61.54**	59.53
DLHN	53.85	**70.77**	**61.54**	**66.15**	60	**64.62**	53.85	58.46	53.85	60	**60.31**

Furthermore, Fig. 7b illustrates the confusion matrix of our proposed DLHN based on four emotion categories. According to Fig. 7a, "negative" and "other" have more samples than "positive" and "surprise". Therefore, the recognition accuracy of "negative" and "other" is higher than the other two categories.



(a) Sample size of each emotion category.

(b) Confusion matrices on combined databases.

Figure 7. micro-expression recognition performance analysis of DLHN per emotion

5. Conclusion 248

In this paper, we proposed a Deep Local-Holistic Network for micro-expression recognition. Specifically, HCRNN is designed to extract local and abundant information from the ROIs related to micro-expression. According to the sparse characteristic of micro-expression, we obtain sparse micro-expression information from original images by RPCA, and utilize RPRNN to extract holistic and sparse features from sparse images. Deep Local-Holistic Network, which fused by HCRNN and RPRNN, captures the local-holistic, sparse-abundant micro-expression information, and boosts the performance of micro-expression recognition. Experimental results on combined databases demonstrate that our proposed method outperforms some state-of-the-art algorithms.

The recognition performance of DLHN remains to be improved due to the limitation of the small sample problem and unbalanced sample distribution. In future work, we will further investigate unsupervised learning as well as data augmentation methods to improve the performance of micro-expression recognition.

Author Contributions: Conceptualization, methodology,and validation, J. Li.,T. Wang., S-J. Wang.; writing—original draft preparation, J. Li.,T. Wang.; writing—review and editing, J. Li., S-J. Wang.; supervision, S-J.Wang. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported in part by grants from the National Natural Science Foundation of China (U19B2032, 62106256, 62061136001), and in part by grants from the China Postdoctoral Science Foundation (2020M680738).

Data Availability Statement: The CASME I database is available at http://fu.psych.ac.cn/CASME/casme-en.php. The CASME II database is available at http://fu.psych.ac.cn/CASME/casme2-en.php. The CAS(ME)^2 database is available at http://fu.psych.ac.cn/CASME/cas(me)2-en.php. The SAMM database is available at http://www2.docm.mmu.ac.uk/STAFF/M.Yap/dataset.php.

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

References

1. Ekman, P.; Friesen, W.V. Nonverbal leakage and clues to deception. *Psychiatry* **1969**, *32*, 88–106.
2. Frank, M.; Herbasz, M.; Sinuk, K.; Keller, A.; Nolan, C. I see how you feel: Training laypeople and professionals to recognize fleeting emotions. In Proceedings of the The Annual Meeting of the International Communication Association. Sheraton New York, New York City, 2009.
3. O'Sullivan, M.; Frank, M.G.; Hurley, C.M.; Tiwana, J. Police lie detection accuracy: The effect of lie scenario. *Law and Human Behavior* **2009**, *33*, 530.
4. Ekman, P. Facial action coding system **1977**.
5. Ekman, P. MicroExpression Training Tool (METT). University of California, San Francisco, 2002.
6. Yan, W.J.; Wu, Q.; Liu, Y.J.; Wang, S.J.; Fu, X. CASME database: a dataset of spontaneous micro-expressions collected from neutralized faces. In Proceedings of the 2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG). IEEE, 2013, pp. 1–7.
7. Yan, W.J.; Li, X.; Wang, S.J.; Zhao, G.; Liu, Y.J.; Chen, Y.H.; Fu, X. CASME II: An improved spontaneous micro-expression database and the baseline evaluation. *PloS one* **2014**, *9*, e86041.
8. Qu, F.; Wang, S.J.; Yan, W.J.; Li, H.; Wu, S.; Fu, X. CAS(ME)^2: A Database for Spontaneous Macro-Expression and Micro-Expression Spotting and Recognition. *IEEE Transactions on Affective Computing* **2017**, *9*, 424–436. doi:10.1109/TAFFC.2017.2654440.
9. Davison, A.K.; Lansley, C.; Costen, N.; Tan, K.; Yap, M.H. SAMM: A Spontaneous Micro-Facial Movement Dataset. *IEEE Transactions on Affective Computing* **2018**, *9*, 116–129. doi:10.1109/TAFFC.2016.2573832.
10. Zhijie.; Cheng.; Tim.; Chuk.; William.; Hayward.; Antoni.; Chan.; Janet.; Hsiao. Global and Local Priming Evoke Different Face Processing Strategies: Evidence From An Eye Movement Study. *Journal of Vision* **2015**.
11. Polikovsky, S.; Kameda, Y.; Ohta, Y. Facial micro-expressions recognition using high speed camera and 3D-gradient descriptor **2009**.
12. Pfister, T.; Li, X.; Zhao, G.; Pietikäinen, M. Recognising spontaneous facial micro-expressions. In Proceedings of the 2011 international conference on computer vision. IEEE, 2011, pp. 1449–1456.
13. Zhao, G.; Pietikainen, M. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Transactions on Pattern Analysis & Machine Intelligence* **2007**, pp. 915–928.
14. Wang, S.J.; Yan, W.J.; Zhao, G.; Fu, X.; Zhou, C.G. Micro-expression recognition using robust principal component analysis and local spatiotemporal directional features. In Proceedings of the European Conference on Computer Vision. Springer, 2014, pp. 325–338.
15. Wright, J.; Ganesh, A.; Rao, S.; Peng, Y.; Ma, Y. Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization. In Proceedings of the Advances in neural information processing systems, 2009, pp. 2080–2088.
16. Wang, S.J.; Chen, H.L.; Yan, W.J.; Chen, Y.H.; Fu, X. Face recognition and micro-expression recognition based on discriminant tensor subspace analysis plus extreme learning machine. *Neural processing letters* **2014**, *39*, 25–43.
17. Wang, S.J.; Yan, W.J.; Li, X.; Zhao, G.; Fu, X. Micro-expression recognition using dynamic textures on tensor independent color space. In Proceedings of the 2014 22nd International Conference on Pattern Recognition. IEEE, 2014, pp. 4678–4683.
18. Huang, X.; Wang, S.J.; Zhao, G.; Piteikainen, M. Facial micro-expression recognition using spatiotemporal local binary pattern with integral projection. In Proceedings of the Proceedings of the IEEE international conference on computer vision workshops, 2015, pp. 1–9.
19. Liu, Y.J.; Zhang, J.K.; Yan, W.J.; Wang, S.J.; Zhao, G.; Fu, X. A main directional mean optical flow feature for spontaneous micro-expression recognition. *IEEE Transactions on Affective Computing* **2015**, *7*, 299–310.

20. Huang, X.; Zhao, G.; Hong, X.; Zheng, W.; Pietikäinen, M. Spontaneous facial micro-expression analysis using spatiotemporal completed local quantized patterns. *Neurocomputing* **2016**, *175*, 564–578.
21. Xu, F.; Zhang, J.; Wang, J.Z. Microexpression identification and categorization using a facial dynamics map. *IEEE Transactions on Affective Computing* **2017**, *8*, 254–267.
22. Wang, S.J.; Wu, S.; Qian, X.; Li, J.; Fu, X. A main directional maximal difference analysis for spotting facial movements from long-term videos. *Neurocomputing* **2017**, *230*, 382–389.
23. Patel, D.; Hong, X.; Zhao, G. Selective deep features for micro-expression recognition. In Proceedings of the 2016 23rd International Conference on Pattern Recognition (ICPR). IEEE, 2016, pp. 2258–2263.
24. Wang, S.J.; Li, B.J.; Liu, Y.J.; Yan, W.J.; Ou, X.; Huang, X.; Xu, F.; Fu, X. Micro-expression recognition with small sample size by transferring long-term convolutional neural network. *Neurocomputing* **2018**, *312*, 251–262.
25. Xia, Z.; Peng, W.; Khor, H.Q.; Feng, X.; Zhao, G. Revealing the invisible with model and data shrinking for composite-database micro-expression recognition. *IEEE Transactions on Image Processing* **2020**, *29*, 8590–8605.
26. Li, Y.; Huang, X.; Zhao, G. Joint Local and Global Information Learning With Single Apex Frame Detection for Micro-Expression Recognition. *IEEE Transactions on Image Processing* **2020**, *30*, 249–263.
27. Zhou, L.; Mao, Q.; Huang, X.; Zhang, F.; Zhang, Z. Feature refinement: An expression-specific feature learning and fusion method for micro-expression recognition. *Pattern Recognition* **2022**, *122*, 108275.
28. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P.; et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* **1998**, *86*, 2278–2324.
29. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural computation* **1997**, *9*, 1735–1780.
30. Schuster, M.; Paliwal, K.K. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing* **1997**, *45*, 2673–2681.
31. Graves, A.; Schmidhuber, J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural networks* **2005**, *18*, 602–610.
32. Donoho, D.L.; et al. High-dimensional data analysis: The curses and blessings of dimensionality. *AMS math challenges lecture* **2000**, *1*, 32.
33. Zhang, K.; Huang, Y.; Du, Y.; wang, l. Facial Expression Recognition Based on Deep Evolutional Spatial-Temporal Networks. *IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society*, pp. 1–1.
34. Lin, Z.; Chen, M.; Ma, Y. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. *arXiv preprint arXiv:1009.5055* **2010**.
35. Davison, A.K.; Merghani, W.; Yap, M.H. Objective Classes for Micro-Facial Expression Recognition(submitted). *Royal Society open science*.
36. Li, X.; Hong, X.; Moilanen, A.; Huang, X.; Pfister, T.; Zhao, G.; Pietikäinen, M. Towards reading hidden emotions: A comparative study of spontaneous micro-expression spotting and recognition methods. *IEEE transactions on affective computing* **2017**, *9*, 563–577.
37. Wang, Y.; See, J.; Phan, R.C.W.; Oh, Y.H. LBP with six intersection points: Reducing redundant information in lbp-top for micro-expression recognition. In Proceedings of the Asian Conference on Computer Vision. Springer, 2014, pp. 525–537.
38. Liong, S.T.; Gan, Y.S.; See, J.; Khor, H.Q.; Huang, Y.C. Shallow triple stream three-dimensional cnn (STSTNet) for micro-expression recognition. In Proceedings of the 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019). IEEE, 2019, pp. 1–5.