applied sciences

MDPI

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, Beijing 100101, China; lijt@psych.ac.cn
[2] Department of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China; tingtrip@163.com
[3] Department of Psychology, University of the Chinese Academy of Sciences, Beijing 100049, China
* Correspondence: wangsujing@psych.ac.cn

Abstract: A 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 consisting of the 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**, *12*, 4643. https://doi.org/10.3390/app12094643

Academic Editor: Antonio Fernández-Caballero

Received: 30 March 2022
Accepted: 3 May 2022
Published: 5 May 2022

1. Introduction

With the explosive growth of multimedia materials and the rapid advancement of deep learning, the technology of face recognition [1–3] and intelligent expression analysis [4–6] is becoming more and more developed. They have been applied in many fields, such as the face unlocking of intelligent devices, human–computer interaction, and face-based emotion understanding. For instance, El Morabit et al. proposed an off-the-shelf CNN architectures to perform an automatic pain estimation from facial expressions [4].

Meantime, as an important nonverbal cue for emotional understanding, facial micro-expression (micro-expression) is an involuntary and momentary facial expression, with a brief duration of less than 500 ms [7]. 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, as illustrated in Figure 1, 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 [8] and national security [9].

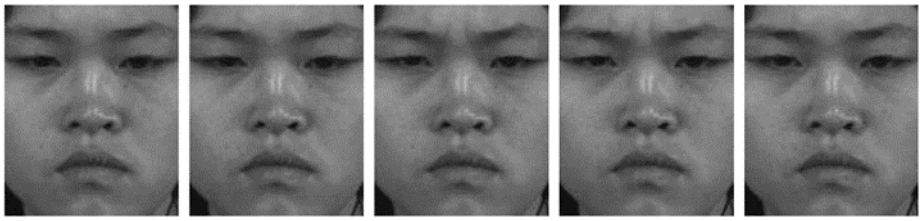


Figure 1. Micro-expression sample with tense emotion: there is a slight inward tucking at the brow, and the movement duration is less than 300 milliseconds (ms). (Sample from CASME II dataset [10]).

Micro-expression is difficult to detect through the naked eye and requires a trained professional to recognize [8]. In order to help people recognize micro-expression, Ekman et al. developed the Facial Action Coding System (FACS) [11] and defined the muscle activity of facial expressions as action units (AU). Meanwhile, they also developed the micro-expression Training Tool (micro-expressionTT) [12]. 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 four spontaneous micro-expression databases: CASME I [13], CASME II [10], CAS(ME)^2 [14] and CAS(ME)^3 [15]. In 2016, Davison et al. released the Spontaneous Actions and Micro-Movements (SAMM) [16] dataset with demographic 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 in 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 and 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 Figure 2, the process of individual cognitive micro-expressions moves from global cognition to local-focused attention and finally to global decision making [17]. Inspired by this theory, we try to algorithmically enhance the ability of the network to learn features with a limited sample size. Thus, 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, and (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. In sum, our proposed DLHN network improves the performance of micro-expression recognition by comprehensively extracting micro-expression spatio-temporal features from both local detail and global sparsity levels.

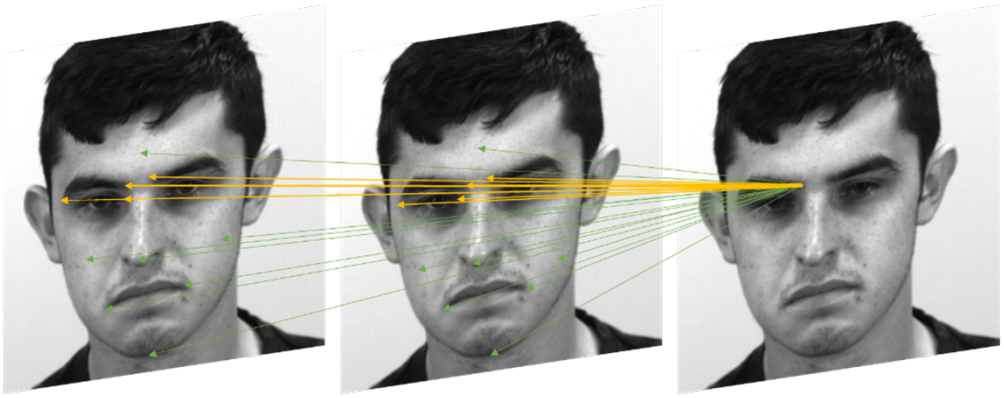


Figure 2. Global (green clipping head) and local area of interest (yellow arrow) tracking of micro-expression action. (Sample from SAMM dataset [16]).

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. [18] divided the face into specific regions and recognized the motion in each region using a 3D-Gradients orientation histogram descriptor. Tomas Pfister et al. [19] designed the first spontaneous micro-expression database (SMIC) and used LBP-TOP [20] to extract dynamic and appearance features of micro-expressions. Wang et al. [21] adopted robust Principal Component Analysis (RPCA) [22] to extract sparse micro-expression information and Local Spatiotemporal Directional Features. Wang et al. introduced a discriminant tensor subspace analysis (DTSA) [23] to preserve the spatial structure information of micro-expression images. Furthermore, they treated a micro-expression video clip as a fourth-order tensor and transformed the color information from RGB into TICS to improve the performance [24]. Huang et al. [25] show a spatiotemporal facial representation to characterize facial movements and used LBP to extract appearance and motion features. Liu et al. [26] proposed a simple, effective Main Directional Mean Optical-flow features (MDMO) and adopted an SVM classifier to recognize micro-expression. Huang et al. [27] 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. [28] proposed a Facial Dynamics Map method by depicting micro-expression characteristics in different granularities. Wang et al. [29] 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. Furthermore, addressing the SSS problem, Li et al. [30] performed data augmentation based on their proposed local temporal pattern for micro-expression analysis.

Recently, the outstanding performance of deep learning has attracted the attention of many researchers to develop micro-expression recognition algorithms. Patel et al. [31] used the pre-trained ImageNet-VGG-f CNN to extract features of each frame in micro-expression

video clips. Wang et al. [32] 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. [33] investigated a low-complexity recurrent convolutional neural network (RCN) based on cross-database micro-expression recognition. Li et al. [34] performed a joint local and global information learning on apex frame for micro-expression recognition. Zhou et al. [35] 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

A 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 it is designed to process still images. It has been shown to outperform many other techniques [36].

As introduced in Section 1, the handcrafted 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 [37], which uses memory cells with multiplicative gate units to process information. It has been shown to outperform RNN in learning long sequences.

In addition, RNN only takes into account the past context. To solve the problems, a bidirectional RNN (BRNN) is created [38], which can process data in both past and future information. Subsequently, Graves et al. [39] proposed a bidirectional LSTM (BLSTM), which has better performance than LSTM in processing long contextual information of complex temporal dynamics.

Since micro-expressions are very subtle, it is not 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

Currently, there are very many mature techniques for signal and image processing that can perform the denoising of images fed into deep learning networks, such as wavelet [40] and compressive sensing [41–43]. In the previous study [44], we tried to retain the principal components and remove irrelevant information such as noise by PCA. The micro-expression action information was considered as noise and removed, making it impossible to obtain valid experimental results by adapting parameters and other means. In the sample processing of micro-expressions, we cannot remove noise directly in the acquisition process because, for micro-expressions, noise is instead common and obvious facial features, such as face contours that can be used as face identity features.

Donoho et al. [45] demonstrated that the observed data could be separated efficiently and exactly into sparse and low-rank structures in high-dimensional spaces. Then, an ide-

alized "robust principal component analysis" problem is proposed to recover a low-rank matrix **A** from highly corrupted measurements **E**:

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

where **A** is the reserved 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. [24] 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 Figure 3, 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.

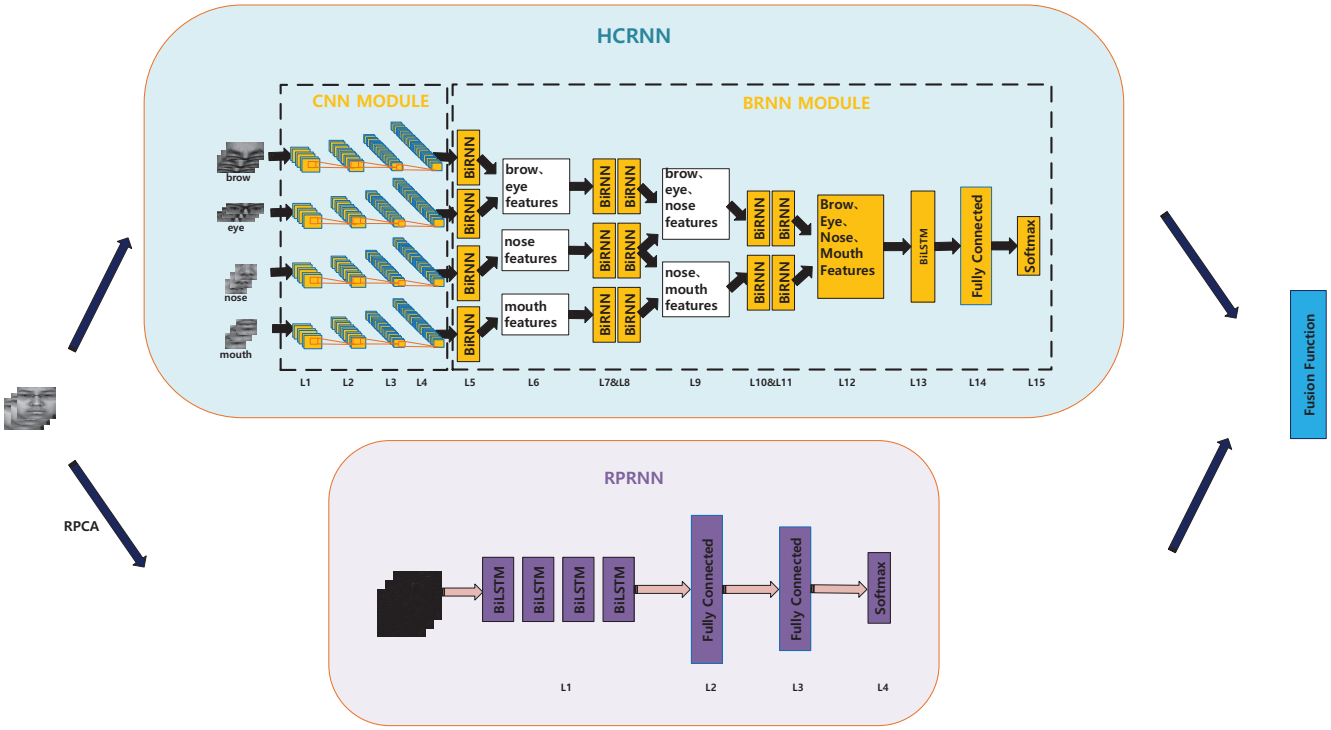


Figure 3. Our proposed Deep Local-Holistic Network. (1) The local network, i.e., HCRNN. The facial micro-expression image is divided into four regions of interest and then fed into four hierarchical CNN modules to extract local-still features. In addition, local dynamic features are learned by the BRNN 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. The activation functions for DLHN are listed in Table 1.

Table 1. Activation function for DLHN.

	Layer	Activation Function
HCRNN	CNN: L1-4	ReLU
	BRNN: L5, 8 and 11	Tanh
	FC L13	Softmax
RPRNN	BiLSTM: L1	ReLU
	FC L2	ReLU
	FC L3	Softmax

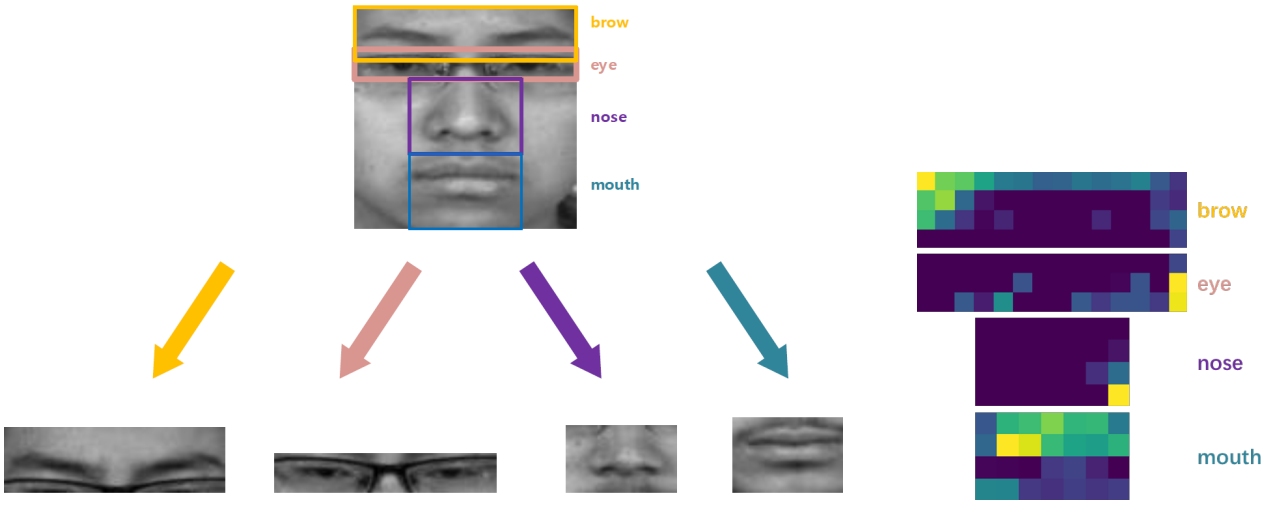
3.1. HCRNN for Local Features

As illustrated in the top block of Figure 3, the HCRNN Model is constructed by the CNN Module and the BRNN Module. First, the CNN Module contains four hierarchical CNNs (HCNNs) to extract local features from ROIs. Then, the 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, and mouth, are used for the local micro-expression feature extraction (Figure 4a). 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 sizes of the eyebrows, eyes, nose, and mouth are 112 × 33, 112 × 20, 56 × 32, and 56 × 38, respectively. Furthermore, considering the integrity of each ROI, the adjacent ROIs may have overlapping portions.

As shown in the HCRNN block of Figure 3, 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 2. 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, Figure 4b presents the feature maps of L4 in HCRNN.



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

Figure 4. ROIs based on eyebrows, eyes, nose and mouth, and the corresponding feature maps of L4 in HCRNN.

Table 2. The HCNN structure.

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

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 [46] is adopted to process temporal variation in micro-expressions.

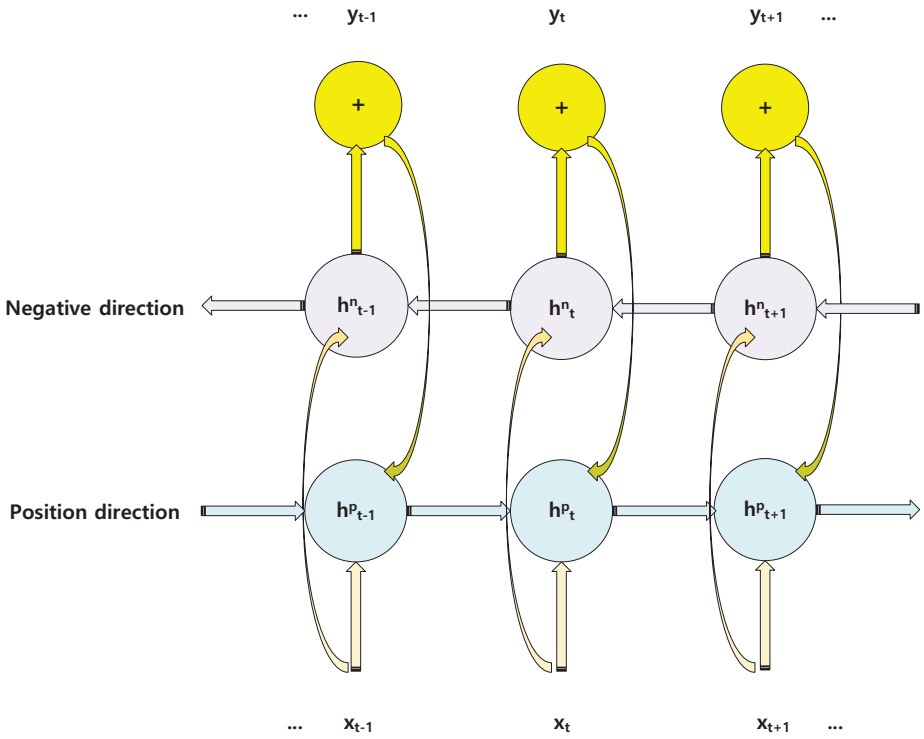


Figure 5. General structure of BRNN. x_t is input data in t time. y_t is output data in t time. h_t^p and h_t^n represent the hidden state in positive and negative directions, respectively.

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 Figure 5). Similar steps of L6 and L7 are repeated in the L8 and L9 layers. A global temporal feature is obtained through the concatenation in the L10 layer

and the BLSTM in the L11 layer. We classify micro-expressions by an FC layer in L12 of HCRNN and obtain probabilistic outputs by the softmax layer in L13 of HCRNN:

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

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

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

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

3.2. RPRNN for Holistic Features

3.2.1. Input: Sparse Micro-Expression Obtained By RPCA

Due to the short duration and low intensity of micro-expression movement, micro-expressions could be considered as sparse data. Hence, RPCA [22] 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 are the height and width in pixels of each frame, respectively, and 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:

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

where A is a low-rank matrix, B is a sparse matrix, $\mathrm{rank}(\cdot)$ 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 the nuclear norm; $\|\cdot\|_1$ denotes the ℓ_1-norm, which counts the sum of all elements in matrix; and λ is a positive weighting parameter ($\lambda > 0$). Lin et al. [47] 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

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 Figure 3. Specifically, 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 the micro-expression is estimated by the softmax layer:

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

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:

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

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

3.3. Model Fusion

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

$$O(x_i) = a\mathrm{P}_{hi}(x_i) + (1-a)\mathrm{P}_{ri}(x_i) \tag{8}$$

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

4. Experiments

4.1. Databases and Validation Protocols

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 3 presents the details of these four databases.

Table 3. 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

- **Emotion category**: The number of emotion classes is different in these four databases, and micro-expression samples are labeled by taking different AU 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. [48]. Finally, we select 650 samples from the combined dataset and divide them into four emotion labels:

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

  Specifically, Negative consists of anger, disgust, sadness, and fear. Figure 6 shows the sample size of each emotion category.
- **Validation Protocol**: Since there are only 650 video samples in the combined database, the sample size for training, validation, and testing would be small with a straightforward division. In order to verify the model performance more fairly and eliminate the effect of individual differences on the results, we adopt a 10-fold cross-validation method. In particular, the samples are equally randomly distributed into 10 folds, of which 9 folds are used for training and the remaining one for testing, and the average of 10 validations represents the accuracy of the model.

- **Evaluation metric**: For the evaluation fo the micro-expression recognition results, a common evaluation criterion is the accuracy of recognition [49], i.e., the proportion of correct predictions among the total number of micro-expression video samples:

$$Accuracy_k = \frac{\sum_i^{nb\_emo} TP_i^k}{nb\_ME_k} \tag{10}$$

$$Accuracy = \frac{\sum_k^{10} Accuracy_k}{10} \tag{11}$$

where k is the index of the test fold in the 10-fold cross-validation; $i \in [1, 2, 3, 4]$, i.e., the emotion label index; nb_emo denotes the number of emotion categories, i.e., 4 in our article; and nb_ME_k represents the total number of micro-expression videos in the kth test fold.

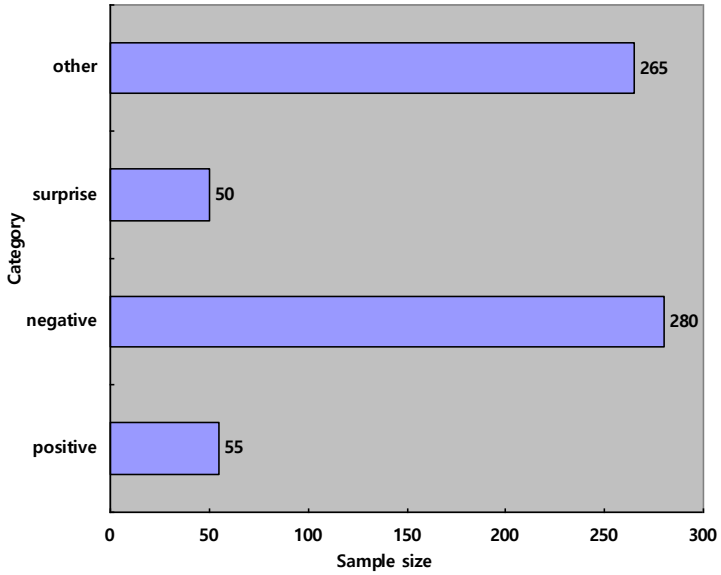


Figure 6. Sample size of each emotion category.

4.2. Preprocessing and Parameter Configuration

4.2.1. Possessing

- **HCRNN**: 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 the CNN module. To guarantee the integrity of each part, ROIs have overlapping areas, and the size of brow, eye, nose, and mouth regions are 112 × 33, 112 × 20, 56 × 32, and 56 × 38, respectively.
- **RPRNN**: The original micro-expression frames are processed by RPCA to obtain the sparse micro-expression images. Figure 7 illustrates an example of micro-expression images processed by RPCA. Then, the sparse images are fed to RPRNN to obtain holistic features.

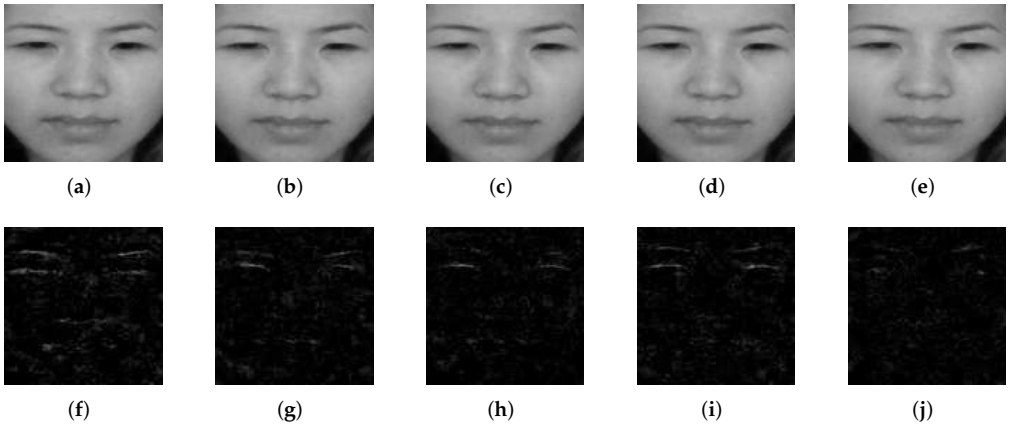
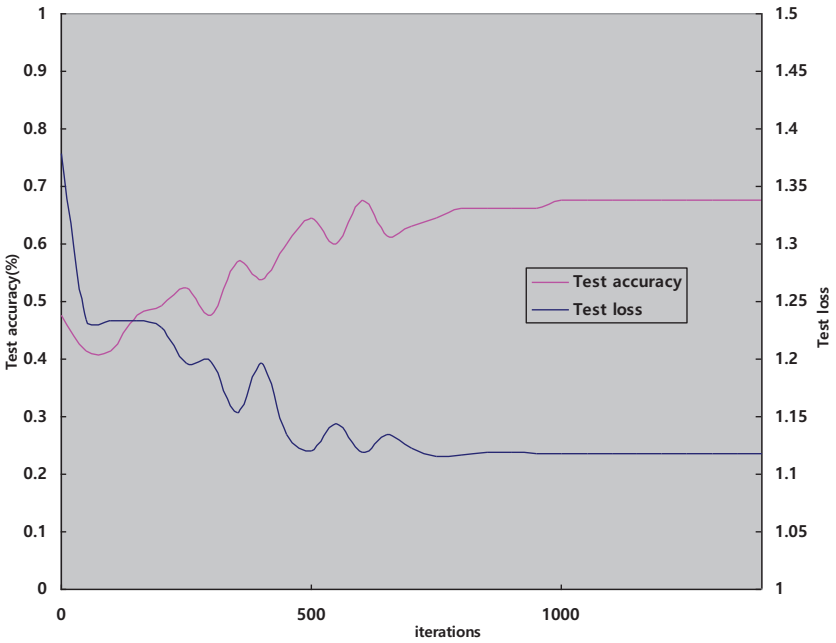


Figure 7. An example of RPCA on micro-expression images: (**a**–**e**) are the original micro-expression images; (**f**–**j**) are the enhanced display for the corresponding extracted sparse information by multiplying each pixel by 2.

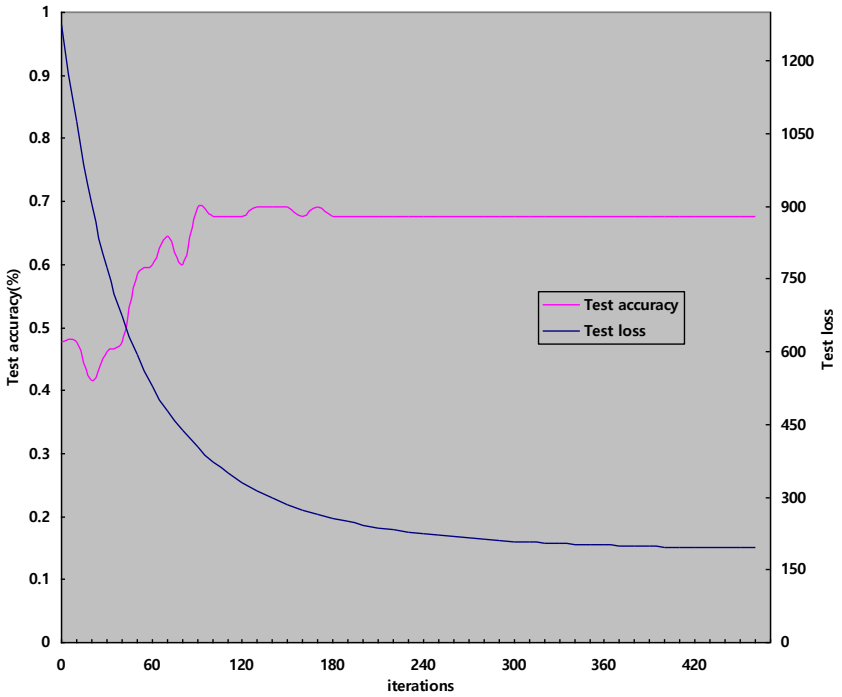
4.2.2. Parameter Configuration for DLHN

- **HCRNN**: The convolution kernel size of the HCNN is set to 3×3, and the size of the pooling kernel is 2×2. The strides of the convolution and pooling layers are set as 1 and 2. In the training stage, the learning rate adopts exponential decay with the initial value equal to 0.85. We update all weights in each iteration with mini-batch samples whose size is 45. The number of epochs is 1500. The iteration curves in Figure 8a represent the trend of the loss and accuracy values in the testing set.
- **RPRNN**: In the model, the attenuation method of the learning rate and the update mode of the weights are the same as the HCRNN, and the value of the learning rate is initialized to 0.01. Same as in the HCRNN, the number of epochs is 1500. In the training stage, we update all weights in each iteration. Figure 8b 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

Figure 8. *Cont.*



(**b**) RPRNN

Figure 8. Network iteration curves.

4.3. Parameter Analysis and Ablation Study

4.3.1. Parameter Analysis

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 4 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.

Table 4. Facial micro-expression recognition accuracy (%) of our proposed DLHN with different parameter a in 10-fold cross-validation dataset. The maximum value in each fold and the maximum mean value are bolded.

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

4.3.2. Ablation Study

To demonstrate the effectiveness of each of our proposed modules as well as combinations, we conduct the following ablation experiments, i.e., data validation on HCRNN,

RPRNN, and the combination of both, DLHN, based on the same experimental setup. The experimental results are listed in Table 5. The mean accuracies of the HCRNN and the RPRNN are 55.08% and 59.53%, respectively. 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 a 60.31% mean accuracy. In addition, RPRNN obtained the best performances in three folds (fold7, fold8, and fold10), which demonstrates the efficiency of the holistic sparse spatio-temporal feature extraction capacity of the RPRNN.

Table 5. Ablation study on the HCRNN and RPRNN modules. The maximum value in each fold and the maximum mean value are bolded.

Method	Fold1	Fold2	Fold3	Fold4	Fold5	Fold6	Fold7	Fold8	Fold9	Fold10	Mean
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**

4.4. DLHN Performance Analysis

4.4.1. Comparison with SOTA Methods

In the choice of comparison methods, among the handcrafted-feature-based methods, we choose the classical FDM features and LBP features [50], as well as the variant of LBP features (LBP-SIP) [51]. 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 [52], RCN(_a,_w,_s, and _f) [33], and Feature Refinement (FR) [35], respectively. Moreover, we all reproduced these methods with the same data configuration.

Table 6 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. Among the deep learning approaches, the Feature Refinement (FR) approach achieved the highest accuracy rate of 56%. Our proposed DLHN method, with global sparse and local detailed spatio-temporal feature extraction, achieves a better performance than all methods, i.e., an accuracy rate of 60.31%.

Table 6. 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. The maximum value in each fold and the maximum mean value are bolded.

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
DLHN	53.85	**70.77**	**61.54**	**66.15**	60	**64.62**	53.85	58.46	53.85	60	**60.31**

Two reasons mainly cause the lack of high performance in micro-expression recognition:

- First, micro-expressions are involuntary and rapidly flowing facial expressions of individuals, which are subtle, brief, and localized. The recognition rate of micro-

expressions in videos by the naked eye is less than 50%, even for professionally trained experts [8]. Similarly, it is very challenging for traditional feature extraction methods and deep learning methods to extract micro-expression features with representational properties. Deep learning networks targeting fine-grained feature learning may be able to improve performance, e.g., by drawing on fine-grained object recognition network designs.

- Second, the small sample size of micro-expressions limits the ability of deep learning to further mine micro-expression features. The maximum sample size of a single database containing micro-expression videos is only 256. In this paper, we combine three common micro-expression databases for analysis, and there are only 650 samples in total. However, the amount of data of this size is still very small compared to face recognition and expression recognition. The performance of micro-expression recognition should be improved in the future when the amount of micro-expression data increases.

4.4.2. Micro-Expression Recognition Per Emotion Class

In addition, to evaluate the algorithm's recognition performance for each emotion class, we analyze it through the confusion matrix. By showing the number of TPs, FPs, TNs, and FNs obtained by the algorithm under different classifications, we can analyze the algorithm's performance in recognizing different emotions. Figure 9 illustrates the confusion matrix of our proposed DLHN based on four emotion classes.

	Positive	Negative	Surprise	Others
Positive (55)	0	11	0	44
Negative (280)	0	208	0	72
Surprise (50)	0	19	0	31
Others (265)	0	81	0	184

Figure 9. Micro-expression recognition performance analysis of DLHN per emotion: confusion matrix on combined databases.

Given that feature learning is very difficult, when an emotion class has more samples, its recognition performance will be relatively stronger. This is because the model is able to learn the corresponding features from more data. According to Figure 6, "Negative" and "other" have more samples than "positive" and "surprise". Thus, the recognition accuracy of "negative" and "other" is higher than the other two emotion classes.

5. Conclusions and Perspective

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 SOTA algorithms. In particular, we achieved an accuracy of 60.31% in recognition with a combination of four micro-expression databases. In comparison with other SOTA methods listed in Table 4 (in the manuscript), our method

outperforms not only the traditional handcrafted feature extraction methods but also the recently published deep learning-based micro-expression recognition methods.

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.L., T.W. and S.-J.W.; writing—original draft preparation, J.L. and T.W.; writing—review and editing, J.L. and S.-J.W.; supervision, S.-J.W. 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), in part by grants from the China Postdoctoral Science Foundation (2020M680738), and in part by Open Research Fund of the Public Security Behavioral Science Laboratory, People's Public Security University of China (2020SYS12).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Adjabi, I.; Ouahabi, A.; Benzaoui, A.; Taleb-Ahmed, A. Past, present, and future of face recognition: A review. *Electronics* **2020**, *9*, 1188. [CrossRef]
2. Adjabi, I.; Ouahabi, A.; Benzaoui, A.; Jacques, S. Multi-block color-binarized statistical images for single-sample face recognition. *Sensors* **2021**, *21*, 728. [CrossRef]
3. Khaldi, Y.; Benzaoui, A.; Ouahabi, A.; Jacques, S.; Taleb-Ahmed, A. Ear recognition based on deep unsupervised active learning. *IEEE Sens. J.* **2021**, *21*, 20704–20713. [CrossRef]
4. El Morabit, S.; Rivenq, A.; Zighem, M.E.n.; Hadid, A.; Ouahabi, A.; Taleb-Ahmed, A. Automatic pain estimation from facial expressions: A comparative analysis using off-the-shelf CNN architectures. *Electronics* **2021**, *10*, 1926. [CrossRef]
5. Perusquia-Hernandez, M.; Hirokawa, M.; Suzuki, K. A wearable device for fast and subtle spontaneous smile recognition. *IEEE Trans. Affect. Comput.* **2017**, *8*, 522–533. [CrossRef]
6. Perusquía-Hernández, M.; Ayabe-Kanamura, S.; Suzuki, K.; Kumano, S. The invisible potential of facial electromyography: A comparison of EMG and computer vision when distinguishing posed from spontaneous smiles. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Glasgow, UK, 4–9 May 2019; pp. 1–9.
7. Ekman, P.; Friesen, W.V. Nonverbal leakage and clues to deception. *Psychiatry* **1969**, *32*, 88–106. [CrossRef]
8. 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 Annual Meeting of the International Communication Association, Sheraton New York, New York City, NY, USA, 7–11 April 2009.
9. O'Sullivan, M.; Frank, M.G.; Hurley, C.M.; Tiwana, J. Police lie detection accuracy: The effect of lie scenario. *Law Hum. Behav.* **2009**, *33*, 530. [CrossRef]
10. 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. [CrossRef]
11. Ekman, P.; Friesen, W.V. Facial action coding system. *Environ. Psychol. Nonverbal Behav.* **1978**. [CrossRef]
12. Paul Ekman Group. *MicroExpression Training Tool (METT)*; University of California: San Francisco, CA, USA, 2002.
13. 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), Shanghai, China, 22–26 April 2013; pp. 1–7.
14. 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 Trans. Affect. Comput.* **2017**, *9*, 424–436. [CrossRef]
15. Li, J.; Dong, Z.; Lu, S.; Wang, S.J.; Yan, W.J.; Ma, Y.; Liu, Y.; Huang, C.; Fu, X. CAS(ME)^3: A Third Generation Facial Spontaneous Micro-Expression Database with Depth Information and High Ecological Validity. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**.
16. Davison, A.K.; Lansley, C.; Costen, N.; Tan, K.; Yap, M.H. SAMM: A Spontaneous Micro-Facial Movement Dataset. *IEEE Trans. Affect. Comput.* **2018**, *9*, 116–129. [CrossRef]
17. Cheng, Z.; Chuk, T.; Hayward, W.; Chan, A.; Hsiao, J. Global and Local Priming Evoke Different Face Processing Strategies: Evidence From An Eye Movement Study. *J. Vis.* **2015**, *15*, 154. [CrossRef]

18. Polikovsky, S.; Kameda, Y.; Ohta, Y. Facial micro-expressions recognition using high speed camera and 3D-gradient descriptor. In Proceedings of the 3rd International Conference on Imaging for Crime Detection and Prevention (ICDP 2009), London, UK, 3 December 2009.
19. Pfister, T.; Li, X.; Zhao, G.; Pietikäinen, M. Recognising spontaneous facial micro-expressions. In Proceedings of the 2011 International Conference on Computer Vision, Washington, DC, USA, 6–13 November 2011; pp. 1449–1456.
20. Zhao, G.; Pietikainen, M. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 915–928. [CrossRef]
21. 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, Zurich, Switzerland, 6–12 September 2014; pp. 325–338.
22. 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 *Advances in Neural Information Processing Systems*; The MIT Press: Cambridge, MA, USA, 2009; pp. 2080–2088.
23. 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 Process. Lett.* **2014**, *39*, 25–43. [CrossRef]
24. 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, Stockholm, Sweden, 24–28 August 2014; pp. 4678–4683.
25. 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 IEEE International Conference on Computer Vision Workshops, Santiago, Chile, 7–13 December 2015; pp. 1–9.
26. 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 Trans. Affect. Comput.* **2015**, *7*, 299–310. [CrossRef]
27. 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. [CrossRef]
28. Xu, F.; Zhang, J.; Wang, J.Z. Microexpression identification and categorization using a facial dynamics map. *IEEE Trans. Affect. Comput.* **2017**, *8*, 254–267. [CrossRef]
29. 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. [CrossRef]
30. Li, J.; Soladie, C.; Seguier, R. Local Temporal Pattern and Data Augmentation for Micro-Expression Spotting. *IEEE Trans. Affect. Comput.* **2020**. [CrossRef]
31. 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), Cancún, Mexico, 4–8 December 2016; pp. 2258–2263.
32. 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. [CrossRef]
33. 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 Trans. Image Process.* **2020**, *29*, 8590–8605. [CrossRef]
34. Li, Y.; Huang, X.; Zhao, G. Joint Local and Global Information Learning With Single Apex Frame Detection for Micro-Expression Recognition. *IEEE Trans. Image Process.* **2020**, *30*, 249–263. [CrossRef]
35. 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 Recognit.* **2022**, *122*, 108275. [CrossRef]
36. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
37. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
38. Schuster, M.; Paliwal, K.K. Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.* **1997**, *45*, 2673–2681. [CrossRef]
39. Graves, A.; Schmidhuber, J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Netw.* **2005**, *18*, 602–610. [CrossRef]
40. Ouahabi, A. *Signal and Image Multiresolution Analysis*; John Wiley & Sons: Hoboken, NJ, USA, 2012.
41. Haneche, H.; Ouahabi, A.; Boudraa, B. New mobile communication system design for Rayleigh environments based on compressed sensing-source coding. *IET Commun.* **2019**, *13*, 2375–2385. [CrossRef]
42. Haneche, H.; Boudraa, B.; Ouahabi, A. A new way to enhance speech signal based on compressed sensing. *Measurement* **2020**, *151*, 107117. [CrossRef]
43. Mahdaoui, A.E.; Ouahabi, A.; Moulay, M.S. Image Denoising Using a Compressive Sensing Approach Based on Regularization Constraints. *Sensors* **2022**, *22*, 2199. [CrossRef] [PubMed]
44. Wang, S.J.; Yan, W.J.; Sun, T.; Zhao, G.; Fu, X. Sparse tensor canonical correlation analysis for micro-expression recognition. *Neurocomputing* **2016**, *214*, 218–232. [CrossRef]
45. Donoho, D.L. High-dimensional data analysis: The curses and blessings of dimensionality. *AMS Math Challenges Lect.* **2000**, *1*, 32.
46. Zhang, K.; Huang, Y.; Du, Y.; Wang, L. Facial Expression Recognition Based on Deep Evolutional Spatial-Temporal Networks. *IEEE Trans. Image Process.* **2017**, *26*, 4193–4203. [CrossRef] [PubMed]

47. Lin, Z.; Chen, M.; Ma, Y. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. *arXiv* **2010**, arXiv:1009.5055.
48. Davison, A.K.; Merghani, W.; Yap, M.H. Objective classes for micro-facial expression recognition. *J. Imaging* **2018**, *4*, 119. [CrossRef]
49. Ben, X.; Ren, Y.; Zhang, J.; Wang, S.J.; Kpalma, K.; Meng, W.; Liu, Y.J. Video-based facial micro-expression analysis: A survey of datasets, features and algorithms. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**. [CrossRef]
50. 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 Trans. Affect. Comput.* **2017**, *9*, 563–577. [CrossRef]
51. 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, Singapore, 1–5 November 2014; pp. 525–537.
52. 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), Lille, France, 14–18 May 2019; pp. 1–5.