Received December 13, 2019, accepted December 27, 2019, date of publication December 31, 2019, date of current version January 8, 2020.

Digital Object Identifier 10.1109/ACCESS.2019.2963201

Facial Expression Recognition by Jointly Partial Image and Deep Metric Learning

NAIGONG YU AND DEGUO BAI

Department of Informatics, School of Artificial Intelligence and Automation, Beijing University of Technology, Beijing 100124, China
Beijing Key Laboratory of Computational Intelligence and Intelligent Systems, Beijing University of Technology, Beijing 100124, China
Digital Community Ministry of Education Engineering Research Center, Beijing University of Technology, Beijing 100124, China

Corresponding author: Naigong Yu (yunaigong@bjut.edu.cn)

This work was supported in part by the National Science Foundation of China under Grant 61573029.

ABSTRACT The performance of facial expression recognition (FER) tends to deteriorate due to high intraclass variations and high interclass similarities. To address this problem, an expression recognition model based on a joint partial image and deep metric learning method (PI&DML) is proposed. First, we propose cropping the active units (AU) that are most closely related to the expression to generate a partial image for feature extraction, which is conducive to mitigating the negative impact of the abovementioned problems to some extent. Second, a novel expression metric loss function (EMLF) is suggested to enhance the intraclass similarities and interclass variations. Finally, superior performance is achieved by jointly optimizing the expression metric loss and classification loss. As demonstrated by the visualization results, the proposed EMLF is effective at increasing the distance between various expressions and reducing the distance between the same expressions. The evaluations on three public expression databases have demonstrated that our method is capable of achieving better results than the state-of-the-art methods.

INDEX TERMS Facial expression recognition, deep metric learning, metric loss function, partial images, jointly optimizing, high intraclass variations, high interclass similarities.

I. INTRODUCTION

Afacial expression is considered a major manifestation of human emotion. Therefore, if a machine is capable of accurately recognizing the facial expressions of human beings, it can improve the outcomes of human-computer interaction (HCI). FER has attracted increasing attention due to its widespread applications in HCI systems such as sociable robots, medical treatments, driver fatigue surveillance and so on [1]. The generic FER framework applied in most works can be split into three major parts, which are face detection, facial feature extraction and classification. Among them, the extraction of the most discriminative facial features is viewed as a significant factor in determining model performance, and these features can be roughly classed into two categories, which are human designed and learned features [2].

The human-crafted features primarily refer to local features, such as the SIFT [3], HOG [4], LBP [5], [6], LPQ [7], *etc*. In addition to the abovementioned methods used to extract the 2D features of static images, focusing on the temporal and spatial information in an image sequence method is also proposed, such as using spatiotemporal covariance descriptors (Cov3D) [8], the temporal modeling of the shape (TMS) [9], expressionlets on a spatiotemporal manifold (STM-ExpLet) [10], *etc*. The FER method based on human-crafted features requires additional classifiers for classification, such as K-NN classifier [11], the SVM classifier [12], and the Hidden Markov model [13]. Although this method has been applied in more cases, its features tend to be relatively singular, and they are susceptible to disruptions caused by head pose and illumination changes [14].

The learned features mainly refer to the features extracted using deep learning methods [15]–[18]. In addition to the features being diversified and robust to illumination changes and different head poses, the methods have also achieved remarkable results in recent years. Khorrami *et al*. performed emotion recognition on video data using both CNNs and RNNs [19]. The CNN is employed to extract the image features, and the RNN is used to express temporal information changes. Yang *et al*. suggested a de-expression residual learning method based on the cGAN [20]. Reference [21] took the SIFT features of landmark points as input data and

The associate editor coordinating the review of this manuscript and approving it for publication was Yudong Zhang.

applied them to a well designed DNN model to extract the optimal discriminative features for expression classification.

Despite the excellent performances achieved by these works in some datasets, they rarely focus on issues such as high intraclass variations and high interclass similarities, which could be caused by diverse head poses, illuminations, occlusions, and personal attributes (skin tone, age, gender, ethnicity, *etc*.), and this remains a challenge to applying the FER in the real-world that needs addressing. As illustrated in Fig.1.(a), there are high intraclass variations and high interclass similarities due to different personal attributes or illuminations, and the learned features in the same class are scattered, which makes it difficult to perform classification in an embedded space. To solve this problem, we combined partial images and the expression metric loss function to reduce the intraclass variations and enhance the interclass variations.

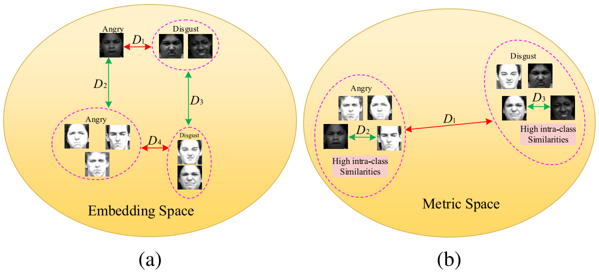


FIGURE 1. **An illustration of features learned by (a) most existing methods (b) the proposed PI&DML model (Note: The length of the green arrow represents the distance between the same expressions, and the length of the red arrow represents the distance between different expressions. The greater the distance, the smaller the similarity.)**

In this paper, distinct from previous works, we have not only researched extracting discriminative features based on the proposed metric loss function, but also explored the importance of partial images in determining FER performance. First, we crop the action units that are most relevant to the expression changes to generate a partial image, which is effective at addressing the abovementioned problems resulting from different illuminations, occlusions and other diverse factors in the overall image. Through experiments, we found that partial images can not only greatly reduce the dimensions of the input data, but also achieve better performance than the original image. Second, we apply the hard sample mining strategy to identify the hardest positive and negative sample pairs in the embedded space, which involves a relatively small amount of computations when compared to the previous metric-based learning. Third, we propose a novel expression metric loss function (EMLF) that is capable of achieving fast convergence to increase not only the similarity between positive sample pairs but also the variations between negative sample pairs, as shown in Fig.1.(b). Finally, by jointly optimizing the expression metric loss and classification loss in a unified framework, an improved classification accuracy compared to the state-of-the-art methods is achieved. Furthermore, we found that the joint optimization of the proposed metric loss and classification loss can use fewer training epoches to achieve faster convergence than a single optimized classification loss.

Overall, the contributions of this work are four-fold. 1) The PI&DML model is proposed which aims to learn discriminative representations with lower intraclass variations and higher interclass distances. 2) A method for constructing partial images is proposed, which is conducive to mitigating the negative impacts of the abovementioned problems to some extent, and can greatly reduce the dimension of the input data and amount of calculations. 3) A novel expression metric loss function (EMLF) is suggested to enhance the intraclass similarities and interclass variations. 4) Superior performance is achieved by jointly optimizing the expression metric loss and classification loss. The evaluations on three public expression databases have demonstrated that our method is capable of achieving better results than the state-of-the-art methods.

The rest of the paper is organised as follows. Section 2 briefly reviews the related topics. Section 3 outlines the methods proposed in this paper, including the method of constructing partial images, the hard sample mining strategy and the proposed expression metric loss function. The experimental results compared with the state-of-the-art methods are given in Section 4. Finally, Section 5 presents a brief conclusion to this paper.

II. RELATED WORK

As mentioned in the introduction, expression recognition methods can be grouped into two categories: still image and sequence-based approaches. Since still image methods are more generic and can also be used to identify expressions from video sequences, we focus on methods for recognizing expressions using still images. Among these, deep learning methods based on convolutional neural network (CNN) architectures have recently shown excellent performance on FER tasks. Despite its popularity, first, the features learned using this method may generate similar representations for different expressions, especially for the same person or the same image brightness. Second, CNNs may generate high variations for the same expression, especially for different people and images with different brightnesses [2]. The emerging deep metric learning methods have demonstrated strong effectiveness in vision tasks with high intraclass variations and high interclass similarities, such as image retrieval [39], [40], person reidentification [41], [42], *etc*., which suggests that deep metric learning can also solve the problems in FER.

Conventional metric learning methods usually learn a linear embedding of the data using the Mahalanobis distance [43], [44], but this is not enough to characterize the nonlinear relationships between sample pairs, which are quite common in real-world applications. Although the kernel trick can be adopted to address this limitation, the expression power of kernel functions is often not flexible enough to capture the nonlinearity in the data [45]. Inspired by deep learning, which can effectively solve the nonlinearity problem of samples, deep metric learning has been proposed to learn nonlinear mappings [46]–[48]. For example, Hu *et al*. proposed a new discriminative deep metric leaning method using deep neural networks for face verification [46], and Wang *et al*. proposed an angular loss for learning better similarity metrics [49]. The multi-similarity loss under the general pair weighting framework was proposed in [28].

For expression recognition, most recently, the island loss function based on the center loss [22] was proposed to reduce the intraclass variations while enlarging the interclass difference [2], which has led to satisfactory performance. Nevertheless, this method is more sensitive to noise samples, and there more hyperparameters that need to be determined. In addition, there are some other research works based on metric learning that have produced positive results [14], [23], [24]. However, a majority of them necessitate the selection of sample pairs and the labeling of identity information in advance, which requires much of extra work.

III. APPROACH

In this section, we will start by presenting the overall framework for the proposed model, and then introduce the approach for constructing the partial image, the hard sample mining strategy and the proposed EMLF.

A. FRAMEWORK

The overall framework of the proposed model is illustrated in Fig.2. First, the mini-batch samples are cropped to generate partial images, which will be sent to the CNN for feature extraction. Then, the hardest positive and negative sample pairs are mined by applying hard sample mining technology in the embedded space for the calculation of the expression loss using the proposed EMLF. Third, the classification loss is calculated at the last fully connected layer. Finally, the overall network is optimized by minimizing the sum of the metric loss and classification loss expressions.

The specific architecture of our proposed model is inspired by the VGG block [50], which consists of a sequence of convolutional layers, followed by a max pooling layer for spatial down sampling. This allows the depth model to be constructed by reusing simple basic blocks. By carefully designing the network parameters, we also found that it is more efficient to use several deep and narrow convolution (i.e. 3×3) layers than a few wide convolution layers. In terms of specific structural parameters, our PI&DML model for both datasets is I(60×30)-C(3,32)-C(3,32)-P(2)-C(3,64)-C(3,64)-P(2)-C(3,128)-C(3,128)-P(2)-FC(512)-FC(256)-FC(128)-FC(n_classes), where I(60×30) means the size of the partial image, and C(3,32) is a convolutional layer with 32 3×3 filters. FC(512) refers to a fully connected layer, with 512 nodes. Additionally, FC(n_classes) is the softmax layer with n_classes outputs, where n_classes represents the number of expression classes for each dataset. P(2) means a 2×2 max pooling layer. The stride of each layer was 1 with the exception of the pooling layer. The value of the stride for each pooling layer was set to 2. Convolutional layers are used to extract the features of expressions, using the hard sample mining strategy and calculating the metric loss in the penultimate layer of the network, more features of the sample are retained, so that the similarity information between the samples can be fully utilized.

B. METHOD OF CONSTRUCTING PARTIAL IMAGE

Human expressions are expressed by the movements of facial components, such as eyes, the mouth and so on. Inspired by this, we select the action units (AU) [25] (eye, nose, mouse) that are considered to be most relevant to the expression to generate a partial image for extracting the discriminative

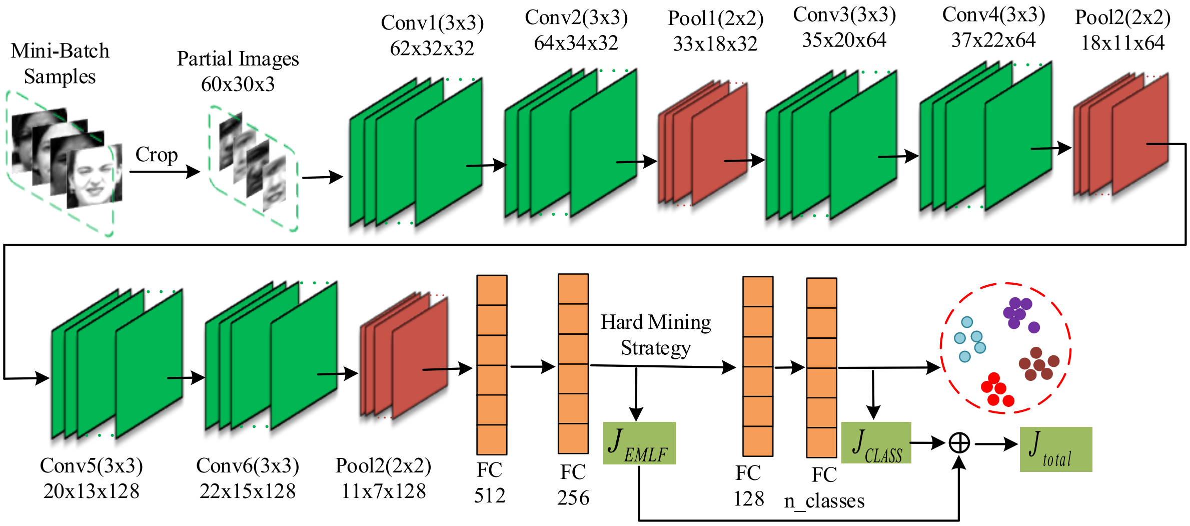


FIGURE 2. The architecture of the proposed PI&DML model. (Note: J_{EMLF} represents expression metric loss, J_{CLASS} represents classification loss and n_classes indicates the number of types of expressions. For the second layer of convolution, 64 × 34 represents the size of the input data, and 32 is the number of convolution kernels. Similarly, network parameters of other layers can be obtained.)

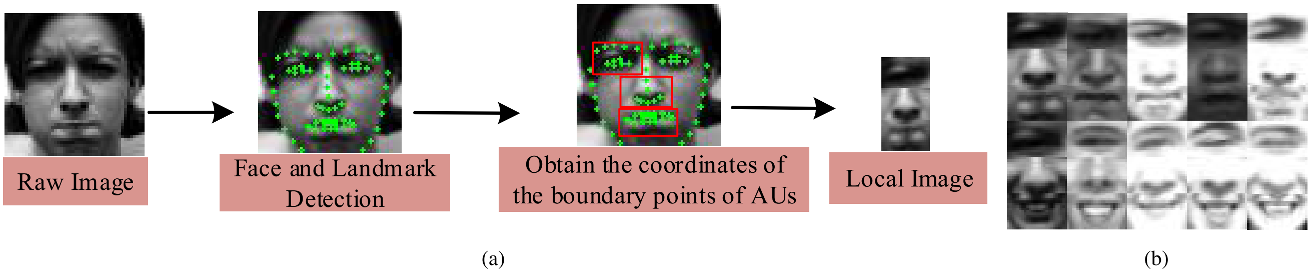

Raw Image
Face and Landmark Detection
Obtain the coordinates of the boundary points of AUs
Local Image
(a)
(b)

FIGURE 3. Partial Image (a) the steps of forming a partial image (b) the display of partial images. The expressions of each line from top to bottom are angry and happy.

features. The steps of forming a partial image are shown in Fig.3.(a), and an example of a partial image is shown in Fig.3.(b). It is clearly indicated that partial images mitigate the influence of personal attributes, illumination and occlusion when compared with the corresponding original images from the CK+ dataset [32] in Fig.4.

FIGURE 4. Original images selected from CK+ dataset, the expressions of each line from top to bottom are angry and happy.

Formally, we apply the face detection method from [26] and the landmark detection method from [27] to obtain face matrix A and the landmark coordinates $(x_i, y_i), i = 0, 1, \cdots 67$. After this step, we can obtain the boundary point coordinates of each AU. According to the boundary point coordinates, each AU can be detected from the original image, and the composition of each AU is shown in (1).

$$\begin{aligned} Eye &= A\left[y_{37} - \tau_1 : y_2 + \tau_1, x_1 - \tau_2 : x_2 + \tau_2\right] \\ Nose &= A\left[y_{28} : y_{33}, x_{31} - \tau_1 : x_{35} + \tau_1\right] \\ Mouse &= A\left[y_{50} - \tau_2 : y_{57} + \tau_1, x_{48} - \tau_2 : x_{54} + \tau_2\right] \end{aligned} \tag{1}$$

where τ_1 and τ_2 represent expanded range at the boundary point of each AU. In order to generate a partial image, each AU needs to be resized to a fixed size S, and the composition of the partial image is indicated in (2), where C denotes the images that are spliced together.

It is known from (1) and (2) that the partial image data are composed of only three parts with respect to the original image data A, which greatly reduces the dimensions of the input data, thereby reducing the amount of calculations.

$$P_{image} = C\{S(Eye), S(Nose), S(Mouth)\} \tag{2}$$

C. HARD SAMPLE MINING STRATEGY

In the embedded space, let $x_i \in \mathbb{R}^d$ be the i_{th} feature of the sample; then, we can obtain a feature matrix $X \in \mathbb{R}^{m\times d}$ for the mini-batch samples, where m indicates the batch size. The similarity between two samples is defined as $S_{ij} =< x_i, x_j >$, where $< \cdot, \cdot >$ denotes the dot product. Then we can obtain an $m \times m$ similarity matrix S, the element of which at (i, j) is S_{ij} for the mini-batch samples. Our aim is to enhance the similarity between the samples of the same classes while reducing the similarity between different classes of samples. It is a simple and easy way to identify samples of the same kind with low similarity (hardest positive pairs) to the current sample (anchor) and, to increase their similarity to the anchor, or to identify different kinds of samples (hardest negative pairs) with higher similarity to the anchor, and to reduce its similarity to the anchor. We apply the hard mining strategy method from [28] to identify the hardest positive pairs and negative pairs as illustrated in Fig.5.

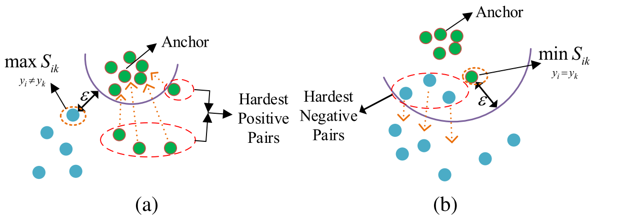


FIGURE 5. Illustration for strategy of hard sample mining. (a) hardest positive pairs mining method (b) hardest negative pairs mining method. (Note: Green dots represent positive samples and blue dots represent samples of different classes than positive samples. The distance between the points represents the degree of similarity, and the further away the samples are, the smaller the similarity is.)

Formally, if x_i is an anchor, the hardest negative pair $\left\{x_i, x_j\right\}$ is selected. If S_{ij} satisfies the condition:

$$S_{ij}^- > \min_{y_k = y_i} S_{ik} - \varepsilon \tag{3}$$

where ε indicates a given margin, and y_k denotes the label of the k_{th} sample. Also, hardest positive pair $\left\{x_i, x_j\right\}$ needs to be met:

$$S_{ij}^+ > \max_{y_k \neq y_i} S_{ik} + \varepsilon \tag{4}$$

D. EXPRESSION METRIC LOSS FUNCTION

Distinct from the previous works such as the contrastive loss [29], triplet loss [30], lifted structure loss [31], *etc.*, a new pair-based expression metric loss function (EMLF) is proposed that removes the need for hyperparameters and is capable of achieving faster convergence rates. Our EMLF is presented as follows:

$$J_{P\_Loss} = \log \sum_{k \in P_i} (-S_{ik}{}^{+} + \sqrt{(-S_{ik}{}^{+})^2 + 1}) \tag{5}$$

$$J_{N\_Loss} = \log \sum_{k \in N_i} (S_{ik}{}^{-} + \sqrt{(S_{ik}{}^{-})^2 + 1}) \tag{6}$$

$$J_{EMLF} = \frac{1}{m} \sum_{i} (J_{P\_Loss} + J_{N\_Loss}) \tag{7}$$

where P_i and N_i represent the hardest positive pairs and hardest negative pairs, respectively. J_{P_Loss} and J_{N_Loss} denote the hardest positive pairs loss and the hardest negative pairs loss, respectively. It can be seen from (6) that reducing this loss value is equivalent to reducing the similarity between the hardest negative sample pairs, and the same reason can be analyzed in (5).

The softmax loss is used to calculate the classification loss, and the L2-Norm is applied to prevent overfitting. The total loss is defined as follows:

$$J_{total} = -\frac{1}{N} \sum_{i} y_i \log(\frac{e^{f_{y_i}}}{\sum_j e^{f_j}}) + J_{EMLF} + \lambda \sum_{w} \|W\|^2 \tag{8}$$

IV. EXPERIMENTS

In this section, to demonstrate the effectiveness of the proposed method for facial expression recognition, experiments are conducted on the CK+ [32], Oulu-CASIA [33] and MMI [34] public facial expression databases to evaluate the proposed model. Furthermore, in order to demonstrate the effectiveness of the proposed partial image method and EMLF, the PI&DML model is compared with three baseline CNNs, which have same network structure as the PI&DML. They are the following: (1) Original images (the images of the detected face) + Softmax loss + EMLF (OSE), (2) Partial images + Softmax loss (PS), and (3) Partial images + Softmax loss + EMLF (PSE).

A. EXPERIMENTAL DATASETS

CK+ dataset: it contains a total of 327 image sequences collected from 118 subjects, each of which is labeled as one of 7 expressions, i.e. anger, contempt, disgust, fear, happiness, sadness and surprise. Each sequence starts with a neutral face, and reaches the peak in the last frame. Similar to other works [2], [14], the last three frames of each sequence are selected to generate 981 images for the experimental dataset.

Oulu-CASIA dataset: it contains totally 480 image sequences collected from 80 subjects, each of which contains one of 6 expressions, i.e. anger, disgust, fear, happiness, sadness and surprise. Similar to the CK+ database, each sequence starts with a neutral facial expression and ends with the facial expression of each emotion. Following the previous works [2], [14], [16], the last three frames are collected as the peak frames of the labeled expression for each sequence. Thus, the Oulu-CASIA dataset contains 1,440 images for our experiments.

MMI dataset: The MMI database consists of 236 image sequences collected from 31 subjects, each sequence is labeled as one of 6 basic expressions, i.e. anger, disgust, fear, happiness, sadness, and surprise, starting from a neutral expression, through a peak phase in the middle, and back to a neutral face at the end. Similar to other works [2], [14], [20], we selected 208 sequences captured in frontal view and three frames in the middle of each image sequence are collected as peak frames associated with the provided label. Hence, there are a total of 624 images used in our experiments.

Preprocessing: The image resolutions of the CK+ dataset, Oulu-CASIA dataset and MMI dataset are 640×490, 320×320, and 186×185, respectively. In the selection of the size of the partial image, we observed the size of all partial images and took an equilibrium value (60×30) from them as the size of the final partial image, this equilibrium value will not cause obvious distortion of all partial images, and it can reduce the dimensions of the input data compared to the initial size of the images of the dataset. Face alignment is performed in works [2], [14], [19] and [20]; and in [2] and [18], they adjusted the contrast of the image. However, the above two operations are not performed on the images here and excellent results are achieved in our work, suggesting that the partial image is effective.

Training/testing strategy: To demonstrate the effectiveness of the proposed method, similar to other works [2], [20], a 10-fold subject-independent cross-validation is adopted for the evaluations conducted on all datasets, where each dataset is further split into 10 subsets. For each run, the data from 8 subsets are used for training and those from the remaining 2 subsets are used for testing. The results are reported as the average of the 10 runs. The training set and the test set cannot have the same kind of expression of the same person at the same time during each run, because if the same kind of expression of the same person appears in both sets at the same time, the model is likely to learn to determine whether is the same person or not in those images, not to determine whether is the same expression.

B. PARAMETERS SETTINGS

In (1), we empirically set $\tau_1 = 25$, $\tau_2 = 9$, and $S = (40, 25)$ for both datasets. In (3) and (4), ε is set to 0.1. For the metric space learning and classification, the Adam [35] optimizer with a batch size of 8 and a learning rate of 0.0001 are used to train the proposed model, the weights of the convolutional layers and fully connected layers were both initialized randomly using the "xaiver" procedure [51], and the number of training epochs is set to 200.

C. EXPERIMENTS RESULTS

Results on the CK+ dataset: The mean accuracy of the 10-fold cross validation is indicated in Table 1. As revealed by the last three results, better recognition accuracy can be achieved by jointly optimizing the expression metric loss and soft loss compared to a single use soft loss, which shows that the proposed metric loss function plays a

TABLE 1. Average accuracy on the CK+ database for seven expressions classification.

Method	Feature	Acc. (%)
HOG 3D [36]	Dynamic	91.44
Cov3D [8]	Dynamic	92.3
STM-Explet [10]	Dynamic	94.19
IACNN [14]	Static	94.37
IL-CNN [2]	Static	94.35
DTAGN [16]	Static	97.25
DERL [20]	Static	97.30
PPDN [23]	Static	97.3
2B(N+M) Softmax [24]	Static	97.1
IDFERM [38]	Static	98.35
OSE(ours)	Static	**98.67**
PS(ours)	Static	**97.23**
PSE(ours)	Static	**98.83**

positive role. Moreover, the recognition accuracy when using partial images is higher than the accuracy of using the original images, which shows that partial images can not only greatly reduce the amount of calculations, but also help to reduce the adverse effects caused by original images. Upon their comparison, the proposed PI&DML model outperforms the human-crafted feature-based methods and deep learning methods. Table 2 shows the confusion matrix of the PI&DML model on the CK+ dataset, and it can be found that our proposed method performs reasonably well at recognizing all emotions.

TABLE 2. Confusion matrix of the PI&DML model for the CK+ database (%).

	An	Co	Di	Fe	Ha	Sa	Su
An	98.88	0	0	0	0	1.1	0
Co	0	100	0	0	0	0	0
Di	0.31	0	99.38	0	0	0.31	0
Fe	0	0	0	99.5	0	0	0.5
Ha	0	0	0.3	0	99.7	0	0
Sa	1.63	0	0.65	0	0	97.71	0
Su	0	0	0.21	0.87	0	0	98.90

Results on the Oulu-CASIA dataset: Table 3 summarizes the comparison results of the Oulu-CASIA dataset, and our proposed method is indicated to improve the accuracy by 7% compared to the current state of the art methods. In addition, it can be clearly seen that the recognition accuracy is low when experiments are performed on the original image, and it is demonstrated that the partial image is capable of mitigating the influences of the illumination, personal attributes and other factors to some extent. The confusion matrix shown in Table 4 indicates the results and demonstrates that all emotions are accurately recognized.

TABLE 3. Average accuracy on the OULU-CASIA database for six expressions classification.

Method	Feature	Acc. (%)
HOG 3D [36]	Dynamic	70.63
COMPACT	Dynamic	91.67
STM-Explet [10]	Dynamic	74.59
DERL [20]	Static	88.0
IL-CNN [2]	Static	77.29
DTAGN [16]	Static	81.46
PPDN [23]	Static	84.59
FN2EN [17]	Static	87.71
IDFERM [38]	Static	88.25
OSE(ours)	Static	**96.32**
PS(ours)	Static	**93.28**
PSE(ours)	Static	**96.53**

TABLE 4. Confusion matrix of the PI&DML model for the OULU-CASIA database (%).

	An	Di	Fe	Ha	Sa	Su
An	95.9	0.6	0.6	0.9	2.0	0
Di	2.85	96.27	0.65	0	0.21	0
Fe	0.45	0.22	94.51	0.68	0	4.11
Ha	0	0.21	3.5	96.27	0	0
Sa	1.3	0.65	0.92	0	96.83	0
Su	0	0	0.45	0	0.45	99.08

Results on the MMI dataset: Since the MMI dataset contains a small number of samples, it is not large enough to train a deep model. Table 5 reports the average accuracy of 10 runs on the MMI database for recognizing six expressions. It can be clearly seen that the recognition accuracy of the proposed PI&DML model is significantly better than those of all the state-of-the-art methods. As shown in the confusion matrix in Table 6, our algorithm was not successful enough for the fear emotion. In particular, most of the fear emotions were confused with surprise, which is the same as the results of other works [2], [16], [20], but our model has a higher recognition rate for other expressions.

TABLE 5. Average accuracy on the MMI database for six expressions classification.

Method	Feature	Acc. (%)
HOG 3D [36]	Dynamic	60.89
STM-Explet [10]	Dynamic	75.12
DERL [20]	Static	73.23
DTAGN [16]	Static	70.24
IL-CNN [2]	Static	70.67
PPDN [23]	Static	84.59
FN2EN [17]	Static	87.71
IDFERM [38]	Static	81.13
OSE(ours)	Static	**93.57**
PS(ours)	Static	**91.27**
PSE(ours)	Static	**94.66**

TABLE 6. Confusion matrix of the PI&DML model for the MMI database (%).

	An	Di	Fe	Ha	Sa	Su
An	98.24	0.58	0.58	0	0.58	0
Di	3.50	93.56	1.17	0.58	1.16	0
Fe	0	0	79.6	4.61	2.63	13.15
Ha	0	0.87	0	99.12	0	0
Sa	1.75	0	1.75	0	95.90	0.58
Su	0	0	3.50	0	0	96.49

D. VISUALIZATION RESULTS

To further illustrate the effectiveness of the proposed method, we visualized the features learned by the OSE, PS and PSE methods on the CK+ dataset, and these feature vectors are

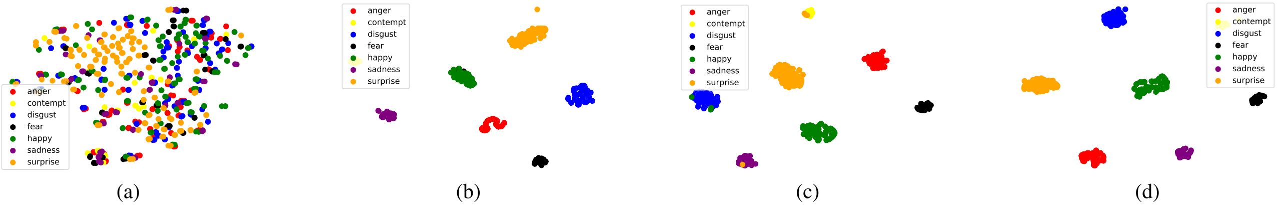


FIGURE 6. Visualization results on CK+ dataset. (a) Visualization of original input data. (b) Visualization of output features learned using the OSE method. (c) Visualization of features learned using the PS method. (d) Visualization of features learned using the PSE method. The data points were automatically grouped by PI&DML model.

visualized using the t-SNE [37], which provides a useful tool for the visualization of the high dimension data. As shown in Fig.6.(a), the input data were spread on a random basis, and most overlap each other.

It can be seen from the comparison between Fig.6.(b) and Fig.6.(d) that the classification effect of the partial images is better than that of the original images when both the classification loss and the proposed metric loss are used, the distance between different classes is relatively large when using partial images, and the features extracted from the last fully connected layer of the proposed model were well separated according to their label.

Comparing Fig.6.(c) with Fig.6.(d), it can be obtained that the classification accuracy can be improved by using the proposed metric loss function, there is almost no overlap between different classes, and the same classes of data can be well clustered together. Therefore, our proposed method reduced the distance between the same classes while increasing the variations between different classes.

E. DISCUSSION ON THE COMPUTATIONAL COST

First, the method of constructing partial images does not require a large amount of calculation, because we only add the step of detecting human eyes, mouth, and nose organs based on face detection, although adding this step slightly reduces the speed of detection, this method simply concatenates the detected partial image together without much calculation. Second, because the partial image is much smaller than the original image, the calculation amount of the model will be reduced in the process of extracting image features. Finally, hard sample mining strategy and metric learning technology did increase the amount of calculation of the model, but by reducing the batch size, the calculation has not increased significantly. For example, in hard sample mining strategy, the similarity matrix $S = X \cdot X^T$, where $X \in \mathbb{R}^{batchsize\times 256}$, $S \in \mathbb{R}^{batchsize\times batchsize}$, so the total number of calculations required to obtain S are: $256 \times 256 \times batchsize \times batchsize$. To reduce the amount of calculations, we choose a smaller batch size of 8. Although the batch size is smaller, the experiments results verify that a better convergence effect can be achieved.

V. CONCLUSION

To address high intraclass variations and high interclass similarities problems in FER, an expression recognition model based on joint partial image and deep metric learning method is proposed in this paper. First, partial image is beneficial to reduce the above problem caused by personal attributes, illuminations, occlusion and other factors to some extent. Second, the proposed EMLF in combination with hard sample mining strategy is applied to learn the nonlinear metric space. Finally, superior performance is achieved by jointly optimizing expression metric loss and classification loss when compared to the state-of-the-art methods on the CK+, Oulu-CASIA and MMI databases.

REFERENCES

[1] S. Li and W. H. Deng, "Deep facial expression recognition: A survey," 2018, *arXiv:1804.08348*. [Online]. Available: https://arxiv.org/abs/1804.08348

[2] J. Cai, Z. Meng, A. S. Khan, Z. Li, J. Oreilly, and Y. Tong, "Island loss for learning discriminative features in facial expression recognition," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Xi'an, China, May 2018, pp. 302–309.

[3] A. Yuce, H. Gao, and J.-P. Thiran, "Discriminant multi-label manifold embedding for facial action unit detection," in *Proc. 11th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, Ljubljana, Slovenia, May 2015, pp. 1–6.

[4] T. Baltrusaitis, M. Mahmoud, and P. Robinson, "Cross-dataset learning and person-specific normalisation for automatic action unit detection," in *Proc. 11th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, Ljubljana, Slovenia, May 2015, pp. 1–6.

[5] S. Moore and R. Bowden, "Local binary patterns for multi-view facial expression recognition," *Comput. Vis. Image Understand.*, vol. 115, no. 4, pp. 541–558, Apr. 2011.

[6] J. Chen, X. Liu, P. Tu, and A. Aragones, "Learning person-specific models for facial expression and action unit recognition," *Pattern Recognit. Lett.*, vol. 34, no. 15, pp. 1964–1970, Nov. 2013.

[7] B. Jiang, B. Martinez, M. F. Valstar, and M. Pantic, "Decision level fusion of domain specific regions for facial action recognition," in *Proc. 22nd Int. Conf. Pattern Recognit.*, Aug. 2014, pp. 1776–1781.

[8] A. Sanin, C. Sanderson, M. T. Harandi, and B. C. Lovell, "Spatio-temporal covariance descriptors for action and gesture recognition," in *Proc. IEEE Workshop Appl. Comput. Vis. (WACV)*, Tampa, FL, USA, Jan. 2013, pp. 103–110.

[9] S. Jain, C. Hu, and J. K. Aggarwal, "Facial expression recognition with temporal modeling of shapes," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCV Workshops)*, Barcelona, Spain, Nov. 2011, pp. 1642–1649.

[10] M. Liu, S. Shan, R. Wang, and X. Chen, "Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 1749–1756.

[11] S. M. Tabatabaei, A. Chalechale, and S. Moghimi, "Facial expression recognition using high order directional derivative local binary patterns," in *Proc. 2nd Int. Conf. Pattern Recognit. Image Anal. (IPRIA)*, Rasht, Iran, Mar. 2015, pp. 1–5.

[12] S. Nigam, R. Singh, and A. K. Misra, "Efficient facial expression recognition using histogram of oriented gradients in wavelet domain," *Multimedia Tools Appl.*, vol. 77, no. 21, pp. 28725–28747, Nov. 2018.

[13] M. H. Siddiqi, R. Ali, A. M. Khan, Y.-T. Park, and S. Lee, ''Human facial expression recognition using stepwise linear discriminant analysis and hidden conditional random fields,'' *IEEE Trans. Image Process.*, vol. 24, no. 4, pp. 1386–1398, Apr. 2015.

[14] Z. Meng, P. Liu, J. Cai, S. Han, and Y. Tong, ''Identity–aware convolutional neural network for facial expression recognition,'' in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Washington, DC, USA, May 2017, pp. 558–565.

[15] F. Zhang, T. Zhang, Q. Mao, and C. Xu, ''Joint pose and expression modeling for facial expression recognition,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake, UT, USA, Jun. 2018, pp. 3359–3368.

[16] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, ''Joint fine–tuning in deep neural networks for facial expression recognition,'' in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, Dec. 2015, pp. 2983–2991.

[17] H. Ding, S. K. Zhou, and R. Chellappa, ''FaceNet2ExpNet: Regularizing a deep face recognition net for expression recognition,'' in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Washington, DC, USA, May 2017, pp. 118–126.

[18] C. Kuo, S. Lai, and M. Sarkis, ''A compact deep learning model for robust facial expression recognition,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Salt Lake, UT, USA, Jun. 2018, pp. 2202–2208.

[19] P. Khorrami, T. Le Paine, K. Brady, C. Dagli, and T. S. Huang, ''How deep neural networks can improve emotion recognition on video data,'' in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Phoenix, AZ, USA, Sep. 2016, pp. 619–623.

[20] H. Yang, U. Ciftci, and L. Yin, ''Facial expression recognition by de-expression residue learning,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake, UT, USA, Jun. 2018, pp. 2168–2177.

[21] T. Zhang, W. Zheng, Z. Cui, Y. Zong, J. Yan, and K. Yan, ''A deep neural network–driven feature learning method for multi-view facial expression recognition,'' *IEEE Trans. Multimedia*, vol. 18, no. 12, pp. 2528–2536, Dec. 2016.

[22] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, ''A discriminative feature learning approach for deep face recognition,'' in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 499–515.

[23] X. Zhao, X. Liang, L. Liu, T. Li, Y. Han, N. Vasconcelos, and S. Yan, ''Peak-piloted deep network for facial expression recognition,'' in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 425–442.

[24] X. Liu, B. V. K. V. Kumar, J. You, and P. Jia, ''Adaptive deep metric learning for identity–aware facial expression recognition,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Honolulu, HI, USA, Jul. 2017, pp. 522–531.

[25] P. Eckman and W. Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Mountain View, CA, USA: Consulting Psychologists Press, 1978.

[26] D. E. King, ''Max-margin object detection,'' 2015, *arXiv:1502.00046*. [Online]. Available: https://arxiv.org/abs/1502.00046

[27] V. Kazemi and J. Sullivan, ''One millisecond face alignment with an ensemble of regression trees,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 1867–1874.

[28] X. Wang, X. Han, W. Huang, D. Dong, and M. R. Scott, ''Multi-similarity loss with general pair weighting for deep metric learning,'' in *Proc. CVPR*, Jun. 2019, pp. 5022–5030.

[29] R. Hadsell, S. Chopra, and Y. Lecun, ''Dimensionality reduction by learning an invariant mapping,'' in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New York, NY, USA, Jul. 2006, pp. 1735–1742.

[30] E. Hoffer and N. Ailon, ''Deep metric learning using triplet network,'' in *Proc. Int. Workshop Similarity-Based Pattern Recognit.*, 2015, pp. 84–92.

[31] H. O. Song, Y. Xiang, S. Jegelka, and S. Savarese, ''Deep metric learning via lifted structured feature embedding,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 4004–4012.

[32] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, ''The extended Cohn–Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression,'' in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, San Francisco, CA, USA, Jun. 2010, pp. 94–101.

[33] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. Pietikäinen, ''Facial expression recognition from near-infrared videos,'' *Image Vis. Comput.*, vol. 29, no. 9, pp. 607–619, Aug. 2011.

[34] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, ''Web–based database for facial expression analysis,'' in *Proc. IEEE Int. Conf. Multimedia Expo*, Amsterdam, The Netherlands, Oct. 2005.

[35] D. P. Kingma and J. Ba, ''Adam: A method for stochastic optimization,'' 2014, *arXiv:1412.6980*. [Online]. Available: https://arxiv.org/abs/1412.6980

[36] A. Klaeser, M. Marszalek, and C. Schmid, ''A spatio–temporal descriptor based on 3D-gradients,'' in *Proc. Brit. Mach. Vis. Conf.*, 2008.

[37] L. van der Maaten and G. Hinton, ''Visualizing data using t-SNE,'' *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

[38] X. Liu, B. V. Kumar, P. Jia, and J. You, ''Hard negative generation for identity-disentangled facial expression recognition,'' *Pattern Recognit.*, vol. 88, pp. 1–12, Apr. 2019.

[39] M. Opitz, G. Waltner, H. Possegger, and H. Bischof, ''BIER—Boosting independent embeddings robustly,'' in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5189–5198.

[40] Y. Duan, W. Zheng, X. Lin, J. Lu, and J. Zhou, ''Deep adversarial metric learning,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake, UT, USA, Jun. 2018, pp. 2780–2789.

[41] W. Chen, X. Chen, J. Zhang, and K. Huang, ''Beyond triplet loss: A deep quadruplet network for person re-identification,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 1320–1329.

[42] J. Zhou, P. Yu, W. Tang, and Y. Wu, ''Efficient online local metric adaptation via negative samples for person re-identification,'' in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2420–2428.

[43] K. Q. Weinberger and L. K. Saul, ''Distance metric learning for large margin nearest neighbor classification,'' *J. Mach. Learn. Res.*, vol. 10, no. 2, pp. 207–244, 2009.

[44] A. Globerson and S. T. Roweis, ''Metric learning by collapsing classes,'' in *Proc. NIPS*, 2006, pp. 451–458.

[45] J. Lu, J. Hu, and J. Zhou, ''Deep metric learning for visual understanding: An overview of recent advances,'' *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 76–84, Nov. 2017.

[46] J. Lu, J. Hu, and Y.-P. Tan, ''Discriminative deep metric learning for face and kinship verification,'' *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4269–4282, Sep. 2017.

[47] H. O. Song, S. Jegelka, V. Rathod, and K. Murphy, ''Deep metric learning via facility location,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 2206–2214.

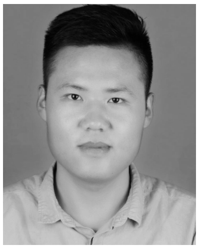
[48] Y. Movshovitz-Attias, A. Toshev, T. K. Leung, S. Ioffe, and S. Singh, ''No fuss distance metric learning using proxies,'' in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 360–368.

[49] J. Wang, F. Zhou, S. Wen, X. Liu, and Y. Lin, ''Deep metric learning with angular loss,'' in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2593–2601.

[50] K. Simonyan and A. Zisserman, ''Very deep convolutional networks for large-scale image recognition,'' in *Proc. ICLR*, 2015.

[51] X. Glorot and Y. Bengio, ''Understanding the difficulty of training deep feedforward neural networks,'' in *Proc. Int. Conf. Artif. Intell. Statist.*, 2010, pp. 249–256.

NAIGONG YU received the bachelor's degree from the Harbin Institute of Technology, in 1989, the master's degree from Shanghai Jiao Tong University, in 1996, and the Ph.D. degree from the Beijing University of Technology, in 2005. He is currently a Professor with the Beijing University of Technology. His research interests include computational intelligence and intelligent systems, robotics, and machine vision.



DEGUO BAI was born in Shandong, China. He received the bachelor's degree from the Shandong University of Technology, in 2018. He is currently pursuing the master's degree with the Beijing University of Technology. His research interests include pattern recognition, computer vision, facial analysis, and intelligent systems.