Facial Representation Extraction by Mutual Information Maximization and Correlation Minimization

Xiaobo Wang[1], Wenyun Sun[2], and Zhong Jin[1(✉)]

[1] School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China
{xiaobowang,zhongjin}@njust.edu.cn

[2] School of Artificial Intelligence, Nanjing University of Information Science and Technology, Nanjing, China
wenyunsun@nuist.edu.cn

Abstract. Facial expression recognition and face recognition are two amusing and practical research orientations in computer vision. Multi-task joint learning can improve each other's performance, which has rarely been studied in the past. This work proposes a joint learning framework to enhance emotional representation and identity representation extraction by incorporating a multi-loss training strategy. Specifically, we propose mutual information loss to ensure that the facial representation is unique and complete and offer correlation loss to extract identity representation using orthogonality constraints. Classification loss is used to learn emotional representation. As a result, we can obtain an unsupervised learning framework to reduce the identity annotation bottleneck using large-scale labelled emotional data for the face verification task. Our algorithm is verified on an artificially synthesized face database: Large-scale Synthesized Facial Expression Dataset (LSFED) and its variants. The identity representation obtained by the algorithm is used for face verification. The performance is comparable to some existing supervised face verification methods.

Keywords: Facial representation extraction · Mutual information · Correlation constraint

1 Introduction

In daily human communication, the information transmitted by face has reached 55% of the total information, plays an essential role in human-computer interaction (HCI), affective computing, and human behaviour analysis. Identity and emotion form the main components in the face domain. To extract discriminative representations, hand-crafted feature operators (i.e., histograms of oriented gradients (HOG), local binary pattern (LBP), and Gabor wavelet coefficients) are used in previous work.

N. T. Nguyen et al. (Eds.): ACIIDS 2021, LNAI 12672, pp. 811–823, 2021.
https://doi.org/10.1007/978-3-030-73280-6_64

However, in recent years, deep learning-based methods [3,15,18,21] are becoming more and more popular and have achieved high recognition accuracy beyond the traditional learning methods. Among them, few jobs consider both identity representation and emotional representation. Li et al. [10] proposed self-constrained multi-task learning combined with spatial fusion to learn expression representations and identity-related information jointly. The novel identity-enhanced network (IDEnNet) can maximally discriminate identity information from expressions. But the network is limited by the identity annotation bottleneck. Yang et al. [24] proposed a cGAN to generate the corresponding neural face image for any input face image. The neural face generated here can be regarded as an implicit identity representation. The emotional information is filtered out and stored in the intermediate layers of the generative model. They use this residual part for facial expression recognition. Sun et al. [17] proposed a pair of Convolutional-Deconvolutional neural networks to learn identity representation and emotional representation. The neutral face is used as the connection point of the two sub-networks, supervising the previous network to extract expression features and input to the latter network to extract identity features.

However, a significant drawback now is that these algorithms require identity supervision labels, among which neutral faces can be regarded as implicit identity labels. It is too strict for facial expression training data. To alleviate such shortcomings, we propose an unsupervised facial orthogonal representation extraction framework. On the premise that only emotional faces and emotion labels are provided, the emotional representation and the identity representation are evaluated using the linear irrelevance of facial attributes. The contributions of this paper are as follows:

- A lightweight convolutional neural network is proposed to extract the identity representation and the emotional representation simultaneously.
- A multi-loss training strategy is proposed, which is a weighted summation of the mutual information loss, the classification loss, and the correlation loss. The mutual information loss measures the relevance between input faces and the deep neural network's output representation. The classification loss is the cross-entropy function commonly used in facial expression recognition tasks. To make up for the lack of identity supervised information, correlation loss is utilized to constrain the linear uncorrelation between the identity representation and the emotional representation.
- The proposed algorithm for facial expression recognition and face verification has achieved outstanding performance on an artificially synthesized face database: Large-scale Synthesized Facial Expression Dataset (LSFED) [17] and its variants [16]. The performance is close to some supervised learning methods.

The rest of this article is organized as follows. Section 2 reviews related work. In Sect. 3, the main methods are proposed. The experiments and results are shown in Sect. 4. Section 5 gives the conclusion.

2 Related Work

2.1 Mutual Information Learning

Representation extraction is a vital and fundamental task in unsupervised learning. The methods based on the INFOMAX optimization principle [4,11] estimate and maximize the mutual information for unsupervised representation learning. They argue that the basic principle of a good representation should be complete and to be able to distinguish the sample from the entire database, that is, to extract the unique information of the sample, for which they introduce mutual information to measure for the first time.

Although mutual information is crucial in data science, mutual information has historically been difficult to calculate, especially for high-dimensional spaces. Mutual Information Neural Estimator (MINE) [2] presents that the estimation of mutual information between high dimensional continuous random variables can be achieved by gradient descent over neural networks. Mutual Information Gradient Estimator (MIGE) [22] argues that directly estimating MI gradient is more appealing for representation learning than estimating MI in itself. The experiments based on Deep INFOMAX (DIM) [4] and Information Bottleneck [1] achieve significant performance improvement in learning proper representation. Some recent works maximize the mutual information between images for zero-shot learning and image retrieval [6,19]. Generally speaking, the existing mutual information-based methods are mainly used to measure the correlation between two random variables, and they are mostly applied to the unsupervised learning of representations.

2.2 Orthogonal Facial Representation Learning

In the absence of identity labels, the algorithm proposed by Sun et al. [17] can obtain relatively clustered identity representation on the facial expression database. In the first half of the network, the neutral face and expression labels are taken as the learning objectives. Then, through the second pair of convolution deconvolution network, the neutral face and expression features are input to reconstruct the original emotional face.

However, in some tasks, the neutral face that belongs to the same person as the original emotional face is difficult to obtain. Excessive training data requirements have become the main disadvantage of the method [17]. Sun et al. [16] put forward an unsupervised orthogonal facial representation learning algorithm. Based on the assumption that there are only two variations in the face space. It should be noted that the emotional representation is invariant to identity change, and the identity representation is invariant to emotion change. To alleviate the dependence on the neutral face, they replace the supervision information with a correlation minimization loss to achieve a similar effect.

Although [16] solves too high database requirements to a certain extent, and the experimental performance on clean databases is also excellent. The reconstruction loss is too strict for facial representation extraction, and much task-independent information is compressed into the middle layer vector. Besides, the

Convolutional-Deconvolutional network will cause excessive expenses. We propose a similar unsupervised facial representation extraction framework to solve these problems, which only uses a lightweight convolutional neural network.

3 Proposed Method

3.1 Deep Neural Network Structure

First, we propose a learning framework consisting of a backbone network and a discriminant network. As shown in Fig. 1, an emotional face is fed into a self-designed VGG-like backbone network to extract identity representation and emotional representation. The network is stacked by 9 basic blocks, and each basic block contains a convolutional layer, a Batch Normalization layer, and an activation layer. The convolutional part consists of nine 3×3 convolutional layers and six pooling layers, and there is no fully connected layer. The convolutional part is formalized as f_θ, where θ represents the trainable parameters of the network. The forward propagation process of the network can be expressed as:

$$(d, l) = f_\theta(x) \tag{1}$$

where x represents an emotional face, d and l represent the corresponding identity representation and emotional representation, respectively. Compared with many complex and deep networks proposed in recent years, our network is simple and sufficient to meet facial expression recognition and face verification tasks. The network's input is 64×64, and the final output is a 519-dimensional global feature vector, where the 512-dimensional vector is identity representation, and the 7-dimensional vector is emotional representation. The specific configuration of the backbone network is shown in Table 1. The discriminant network is designed to estimate mutual information, which will be described in detail in the next section. The three grey squares in Fig. 1 represent three losses, namely mutual information loss L_{mi}, classification loss L_{cls}, and correlation loss L_{corr}.

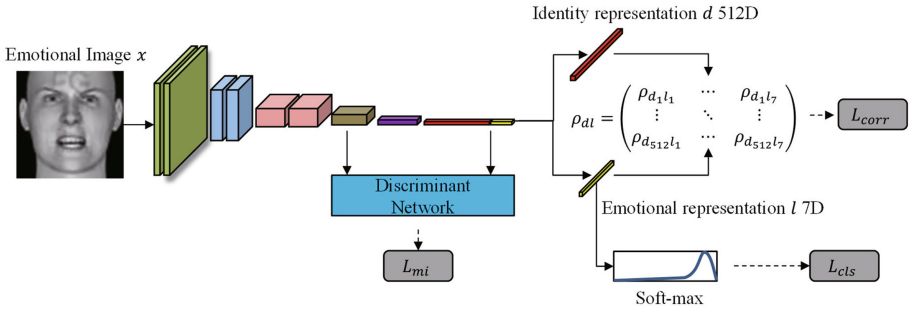

Fig. 1. The overall architecture of the proposed method

Table 1. Structure of the baseline network

Neural network layer	Feature map size	Number of parameters
Input layer	$64 \times 64 \times 1$	0
Convolutional layer	$64 \times 64 \times 16$	192
Convolutional layer	$64 \times 64 \times 16$	2352
Avg pooling	$32 \times 32 \times 16$	0
Convolutional layer	$32 \times 32 \times 32$	4704
Convolutional layer	$32 \times 32 \times 32$	9312
Avg pooling	$16 \times 16 \times 32$	0
Convolutional layer	$16 \times 16 \times 64$	18624
Convolutional layer	$16 \times 16 \times 64$	37056
Avg pooling	$8 \times 8 \times 64$	0
Convolutional layer	$8 \times 8 \times 128$	74112
Avg pooling	$4 \times 4 \times 128$	0
Convolutional layer	$4 \times 4 \times 256$	295680
Avg pooling	$2 \times 2 \times 256$	0
Convolutional layer	$2 \times 2 \times (512 + 7)$	1197333
Avg pooling	$1 \times 1 \times (512 + 7)$	0

A Batch Normalization layer (BN) and a Tanh activation function exist after each convolutional layer. The network uses BN technology to accelerate training and obtain centralized features, facilitating subsequent correlation loss calculations.

3.2 Mutual Information Loss

Previous work has shown that reconstruction is not a necessary condition for adequate representation. The basic principle of a good representation should be complete and to be able to distinguish the sample from the entire database, that is, to extract the unique information of the sample. We use mutual information to measure the correlation of two variables and maximize the correlation measure to restrict that the extracted information is unique to the sample. The overall idea is derived from Deep INFOMAX [4]. X represents the collection of emotional faces, Z represents the collection of encoding vectors and $p(z \mid x)$ represents the distribution of the encoding vectors generated by x, where $x \in X$and $z \in Z$. Then the correlation between X and Z is expressed by mutual information as:

$$I(X, Z) = \iint p(z \mid x)p(x) \log \frac{p(z \mid x)}{p(z)} \mathrm{d}x\mathrm{d}z \tag{2}$$

$$p(z) = \int p(z \mid x)p(x)\mathrm{d}x \tag{3}$$

A useful feature encoding should make mutual information as large as possible:

$$p(z \mid x) = \underset{p(z|x)}{\operatorname{argmax}} I(X, Z) \tag{4}$$

The larger the mutual information means that the $\log \frac{p(z|x)}{p(z)}$ should be as large as possible, which means that $p(z \mid x)$ should be much larger than $p(z)$, that is, for each x, the encoder can find the z that is exclusive to x, so that $p(z \mid x)$ is much greater than the random probability $p(z)$. In this way, we can distinguish the original sample from the database only by z.

Mutual Information Estimation. Given the fundamental limitations of MI estimation, recent work has focused on deriving lower bounds on MI [20,23]. The main idea of them is to maximize this lower bound to estimate MI. The definition of mutual information is slightly changed:

$$\begin{aligned} I(X, Z) &= \iint p(z \mid x)p(x) \log \frac{p(z \mid x)p(x)}{p(z)p(x)} \mathrm{d}x\mathrm{d}z \\ &= KL(p(z \mid x)p(x) \| p(z)p(x)) \end{aligned} \tag{5}$$

To obtain complete and unique facial representation (i.e., identity representation and emotion representation), we maximize the distance between the joint distribution and the marginal distribution to maximize mutual information proposed in Eq. (5). We use JS divergence to measure the difference between the two distributions. According to the local variational inference of f divergence [13], the mutual information of the JS divergence version can be written as:

$$\begin{aligned} JS(p(z \mid x)p(x), p(z)p(x)) = \max_{T} \big(& E_{(x,z)\sim p(z|x)p(x)}[\log \sigma(T(x, z))] \\ & + E_{(x,z)\sim p(z)p(x)}[\log(1 - \sigma(T(x, z)))] \big) \end{aligned} \tag{6}$$

where T is a discriminant network, and σ is the sigmoid function. Refer to the negative sampling estimation in word2vec [9,12,14], x and its corresponding z are regarded as a positive sample pair (i.e., sampled from joint distribution), and x and randomly drawn z are regarded as negative samples (i.e., sampled from marginal distribution). As illustrated in Fig. 2. The discriminant network is trained to score sample pairs so that the score for positive samples is as high as possible, and the score for negative samples is as low as possible. Generally speaking, the right side of Eq. (6) can be regarded as the negative binary cross-entropy loss. For fixed backbone networks, mutual information is estimated (see Eq. (6)). Further, to train the discriminant network and the backbone network at the same time to evaluate and maximize the mutual information, respectively, Eq. (4) is replaced by the following objective:

$$\begin{aligned} p(z \mid x), T(x, z) = \underset{p(z|x), T(x,z)}{\operatorname{argmax}} \big(& E_{(x,z)\sim p(z|x)p(x)}[\log \sigma(T(x, z))] \\ & + E_{(x,z)\sim p(z)p(x)}[\log(1 - \sigma(T(x, z)))] \big) \end{aligned} \tag{7}$$

where $p(z \mid x)$ is the backbone network proposed in Sect. 3.1, $T(x, z)$ is the discriminant network.

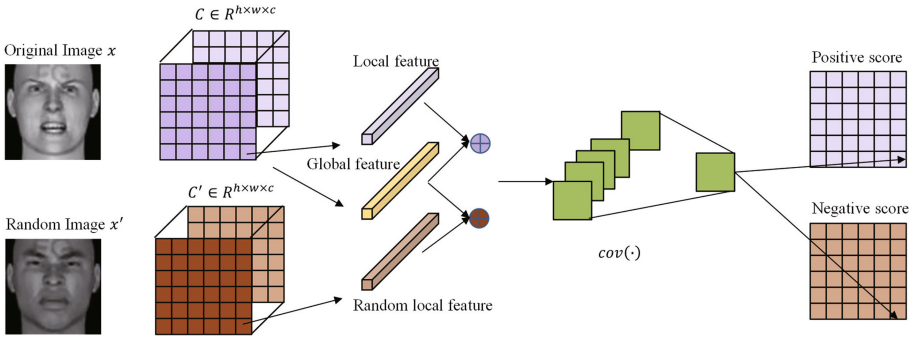

Fig. 2. The forward propagation process of the discriminant network. A random image is selected in a batch, $C \in R^{h\times w\times c}$ is the middle layer feature map. Cov $(\cdot)$ is a 2-layered 1×1 convolutional neural network and $\oplus$ indicates concatenate operation. The global feature is estimated from the original image. Local features and random local features are extracted from the same spatial position.

Mutual Information in a Neural Network. In a neural network, we can compute mutual information between arbitrary intermediate features. Therefore we figure another format of the mutual information in a neural network: $I(f_{\theta_1}(X), f_{\theta_2}(X))$, where f_{θ_1} and f_{θ_2} correspond to activations in different/same layers of the same convolutional network. When f_{θ_1} indicates the input layer and f_{θ_2} represents the top layer of the convolutional network, we call it global mutual information (GMI) because it considers the correlation between the entire faces X and its corresponding global representations Z. However, due to the original face's high dimensionality, it is challenging to directly calculate the mutual information between the network input and output features. And for face verification and facial expression recognition tasks, the correlation of face is more reflected in the local features. Therefore, it is necessary to consider local mutual information (LMI). Let $C \in R^{h\times w\times c}$ denotes the intermediate layer feature map, the mutual information loss is expressed by local mutual information as:

$$L_{mi} = I(C, Z) = \frac{1}{hw} \sum_{i,j} I\left(C_{i,j}, Z\right) \tag{8}$$

where $1 \ll i \ll h$ and $1 \ll j \ll w$. The mutual information between the vector of each spatial position of the feature map and the final global feature vector is calculated. Then the arithmetic mean of them, regarded as the local mutual information, is applied in the representation learning.

3.3 Correlation Loss

The second-order statistics of features has an excellent performance in face tasks and domain adaptation problems. The covariance alignment increases the correlation between the source and target domain by aligning the data distribution of

the source and target domain. On the contrary, to ensure that identity and emotional representations do not interact with each other, we calculate and minimize the pairwise Pearson Correlation Coefficient matrix (PCC) between identity and emotional representations. Compared with covariance, the Pearson Correlation Coefficient is dimensional invariance. It will not lead to a neural network with small weights and small features, which affects the subsequent non-linear feature learning. The Pearson Correlation Coefficient matrix between the identity and emotional representations (See Sect. 3.1, Eq. (1), $d = (d_1, d_2, \cdots, d_{512})$ and $l = (l_1, l_2, \cdots, l_7)$) is aligned to zeros, defined as follows:

$$\rho_{dl} = \begin{pmatrix} \rho_{d_1 l_1} & \cdots & \rho_{d_1 l_7} \\ \vdots & \ddots & \vdots \\ \rho_{d_{512} l_1} & \cdots & \rho_{d_{512} l_7} \end{pmatrix} \tag{9}$$

The Pearson Correlation Coefficients of two random variables d_i and l_j are defined as follows:

$$\rho_{d_i l_j} = \frac{\mathrm{Cov}\left(d_i, l_j\right)}{\sqrt{\mathrm{Var}\left(d_\mathrm{i}\right) \mathrm{Var}\left(l_j\right)}} = \frac{E\left[\left(d_\mathrm{i} - E\left(d_\mathrm{i}\right)\right)\left(l_j - E\left(l_j\right)\right)\right]}{\sigma\left(d_\mathrm{i}\right)\sigma\left(l_j\right)} \tag{10}$$

where $\mathrm{Cov}\left(\cdot\right)$, $\mathrm{Var}\left(\cdot\right)$, $\mathrm{E}\left(\cdot\right)$, $\sigma\left(\cdot\right)$ are functions of covariance, variance, expectation, and standard deviation, respectively. The Eq. (10) shows that the PCC can also be regarded as a normalized covariance, and it varies from -1 to $+1$. -1 means a complete negative correlation, $+1$ means an absolute positive correlation, and zero indicates no correlation. Based on the PCC's properties, we define the correlation loss as follows:

$$L_{\mathrm{corr}} = \sum_{i,j} \left(\rho_{d_i l_j}\right)^2 \tag{11}$$

In a neural network, $\mathrm{E}\left(\cdot\right)$ is always estimated in a mini-batch. Here we propose a fairly simple method to obtain centralized features without additional computation. As shown in Fig. 1, we use the features after Batch Normalization and before the bias addition as the centralized identity representation, and the ground truth emotion label y in the form of C-dimensional one-hot code as the emotional representation. So $y - \frac{1}{C}$ is used as centralized emotional representations. $\mathrm{E}\left(\cdot\right)$ and $\sigma\left(\cdot\right)$ in Eq. (10) can be eliminated to achieve efficient and accurate forward/reverse calculation.

Finally, we use the cross-entropy function to define the expression classification loss L_{cls}:

$$L_{cls} = -\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} y_{i,j} \log \frac{e^{l_{i,j}}}{\sum_{j'=1}^{m} e^{l_{i,j'}}} \tag{12}$$

where n is the mini-batch size, m is the number of expression categories, y is the ground truth label, and l is the predicted probabilities in logarithmic space.

The total loss consists of mutual information loss, correlation loss, and classification loss:

$$L_{\mathrm{total}} = -\alpha L_{mi} + \beta L_{\mathrm{corr}} + L_{cls} \tag{13}$$

Among them, the non-negative α and β balance the importance of the three losses. We will discuss these hyperparameters in detail below.

4 Experiments

4.1 Databases and Preprocessing

To verify the superiority of our proposed algorithm, we used multiple facial expression databases to conduct experiments, including the LSFED, the LSFED-G, the LSFED-GS, and the LSFED-GSB. The generation of the LSFED are based on FaceGen modeller software [5], which strictly follows the definitions of FACS and EMFACS. The LSFED has 105000 aligned facial images. [16] proposed three variants. G represents Gaussian noise with Signal to Noise Ratio (SNR) = 20 dB. S represents random similarity transform. B represents random background patches from the CIFAR-10 database and CIFAR-100 database [7]. The samples in the LSFED, the LSFED-G, the LSFED-GS, and the LSFED-GSB are illustrated in Fig. 3. The four databases are roughly divided into training sets and testing sets with a ratio of 8:2.

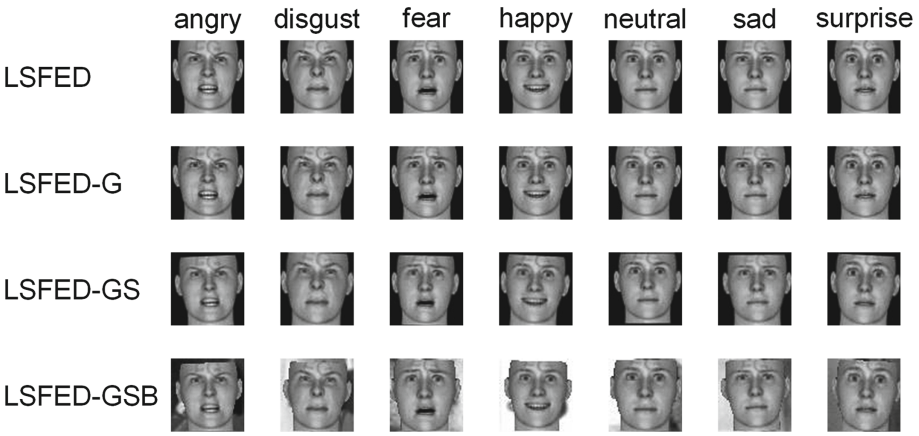

Fig. 3. The samples in the LSFED, the LSFED-G, the LSFED-GS, and the LSFED-GSB

In the training process, we use ADAM optimizer to minimize the total loss in a mini-batch. The experiments were carried out on GeForce GTX 1060. The learning rate is 0.001, and the momentum is 0.8. The whole training process stopped after 100 epochs.

4.2 Experiments Based on the Identity Representation

The learned identity representation is evaluated on a face verification task. We randomly choose 1000 positive pairs (same identity, different expressions) and 1000 negative pairs (different identities, same expression), compute the Euclidean distances between the learned identity representations of these pairs, use the median of the distances as the threshold for face verification. Areas under receiver operating characteristic curves (AUC) and Equal Error Rates (EER) are listed in Table 2 as indicators for evaluating the quality of face verification tasks.

When $\alpha = 0$, $\beta = 0$, there is no external loss to constrain the extracted facial representation except for expression classification loss. The learned identity representation cannot be used for face verification, and none of the 2000 pairs of faces selected randomly is correct. When $\alpha = 1$, $\beta = 0$, we use mutual information loss and facial expression classification loss to learn the identity representation, the face verification performance is better than the original image X on the LSFED-GS and the LSFED-GSB. It verifies the effectiveness of mutual information in extracting face unique information (i.e., emotional information and identity information), as described in Sect. 3.2. When $\alpha = 0$, $\beta = 1$, the identity representation is learned based on the assumption that the identity representation is orthogonal to the emotional representation. When $\alpha = 100$, $\beta = 1$, we obtain the best performance on all four databases, which are 0.999/1.3, 1.000/0.3, 0.983/7.0, and 0.969/8.7.

The proposed method is compared with several existing face verification methods in Table 3. When $\alpha = 100$, $\beta = 1$, the identity representation based face verification outperforms all unsupervised methods and most supervised methods.

Table 2. Experimental performance of face verification

		AUC/EER(%)			
		LSFED	LSFED-G	LSFED-GS	LSFED-GSB
Original image		0.985/6.4	0.985/6.5	0.781/30.1	0.613/42.2
Identity representation	$\alpha = 0$, $\beta = 0$	0.000/100	0.000/100	0.000/100	0.000/99.6
	$\alpha = 0$, $\beta = 1$	0.949/11.3	0.951/10.3	0.826/25	0.692/37.7
	$\alpha = 1$, $\beta = 0$	0.934/13.3	0.956/10.6	0.811/26.9	0.815/26.1
	$\alpha = 1$, $\beta = 1$	0.947/12.8	0.968/10.8	0.790/30.5	0;909/15.8
	$\alpha = 10$, $\beta = 1$	0.996/3.3	0.997/2.6	0.887/20.7	0.779/32.2
	$\alpha = 100$, $\beta = 1$	0.999/1.3	1.000/0.3	0.983/7.0	0.969/8.7

4.3 Experiments Based on the Emotional Representation

We directly use facial expression recognition (FER) accuracy to evaluate the learned emotional representation. When $\alpha = 0$, $\beta = 1$, mutual information loss is suppressed, we can guarantee that emotional representation and identity representation are orthogonal, and this constraint has side effects on facial

Table 3. Comparison with existing face verification algorithms

	Methods	AUC			
		LSFED	LSFED-G	LSFED-GS	LSFED-GSB
Unsupervised	Sun et al.'s unsupervised method [16]	1.000	1.000	0.920	0.768
	Proposed method	0.999	1.000	0.983	0.969
	Original image	0.985	0.985	0.781	0.613
Supervised	2-layered Neural Network	1.000	0.998	0.978	0.970
	AlexNet feature +2-layered NN	0.994	0.991	0.968	0.932
	Sun et al.'s supervised method [17]	1.000	1.000	0.999	0.998

expression recognition, especially the LSFED-GSB database, which decreased from 100% to 92.7%. When $\alpha \geq 1$ and $\beta = 1$, as the ratio of α to β increases (i.e., $\alpha/\beta = 1, 10, 100$), the facial expression accuracy of the LSFED and its variations also increases, which indicates that mutual information loss can assist in extracting more compact expression representation.

Compared with several existing methods, when $\alpha = 100$, $\beta = 1$, the accuracy outperforms the Nearest Neighbor Classifier, the PCA + LDA, and AlexNet [8]. It is comparable to Sun et al.'s methods [16,17] on four facial expression databases (Table 4).

Table 4. Experimental performance of facial expression recognition

Methods	Accuracy (%)			
	LSFED	LSFED-G	LSFED-GS	LSFED-GSB
Sun et al.'s unsupervised method [16]	100.0	100.0	99.9	99.1
Proposed method, $\alpha = 0$, $\beta = 0$	100.0	100.0	100.0	100.0
Proposed method, $\alpha = 0$, $\beta = 1$	99.4	99.3	97.9	92.7
Proposed method, $\alpha = 1$, $\beta = 0$	100.0	100.0	100.0	100.0
Proposed method, $\alpha = 1$, $\beta = 1$	99.7	99.7	98.4	96.5
Proposed method, $\alpha = 10$, $\beta = 1$	99.9	99.9	99.5	98.5
Proposed method, $\alpha = 100$, $\beta = 1$	99.9	99.9	99.5	99.4
Nearest Neighbor Classifier	95.6	94.2	70.8	61.9
PCA + LDA	97.8	99.4	93.1	89.8
Linear SVM	99.8	99.7	85.3	81.9
AlexNet [8]	91.9	97.4	96.3	98.1
Sun et al.'s supervised method [17]	100.0	100.0	99.8	98.7

5 Conclusion

In this paper, we present a novel approach for facial representation extraction (i.e., identity representation and emotional representation), which is based on a lightweight convolutional neural network and a multi-loss training strategy. First, based on the design idea of the VGG network, a lightweight convolutional neural network with only about 1.6 million parameters is proposed. Second, three

losses are proposed to train the network. The mutual information loss is proposed to make sure that the facial representation is unique and complete, and the correlation loss is proposed to leverage orthogonality constraint for identity and emotional representation extraction. The classification loss is used to learn emotional representation. The learning procedure can capture the expressive component and identity component of facial images at the same time. Our proposed method is evaluated on four large scale artificially synthesized face databases. Without exploiting any identity labels, the identity representation extracted by our method is better than some existing unsupervised/supervised methods in the performance of face verification.

Acknowledgments. This work is partially supported by National Natural Science Foundation of China under Grant Nos 61872188, U1713208

References

1. Alemi, A.A., Fischer, I., Dillon, J.V., Murphy, K.: Deep variational information bottleneck. arXiv preprint arXiv:1612.00410 (2016)
2. Belghazi, M.I., et al.: MINE: mutual information neural estimation. arXiv preprint arXiv:1801.04062 (2018)
3. Cai, J., Meng, Z., Khan, A.S., Li, Z., O'Reilly, J., Tong, Y.: Island loss for learning discriminative features in facial expression recognition. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2018, pp. 302–309. IEEE (2018)
4. Hjelm, R.D., et al.: Learning deep representations by mutual information estimation and maximization. arXiv preprint arXiv:1808.06670 (2018)
5. Singular Inversions: FaceGen Modeller (version 3.3) [computer software]. Singular Inversions, Toronto, ON (2008)
6. Kemertas, M., Pishdad, L., Derpanis, K.G., Fazly, A.: RankMI: a mutual information maximizing ranking loss. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14362–14371 (2020)
7. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
8. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. Commun. ACM **60**(6), 84–90 (2017)
9. Le, Q., Mikolov, T.: Distributed representations of sentences and documents. In: International Conference on Machine Learning, pp. 1188–1196 (2014)
10. Li, Y., et al.: Identity-enhanced network for facial expression recognition. In: Jawahar, C.V., Li, H., Mori, G., Schindler, K. (eds.) ACCV 2018. LNCS, vol. 11364, pp. 534–550. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-20870-7_33
11. Linsker, R.: Self-organization in a perceptual network. Computer **21**(3), 105–117 (1988)
12. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
13. Nowozin, S., Cseke, B., Tomioka, R.: f-GAN: training generative neural samplers using variational divergence minimization. In: Advances in Neural Information Processing Systems, pp. 271–279 (2016)
14. Rong, X.: word2vec parameter learning explained. arXiv preprint arXiv:1411.2738 (2014)

15. Shi, Y., Yu, X., Sohn, K., Chandraker, M., Jain, A.K.: Towards universal representation learning for deep face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6817–6826 (2020)
16. Sun, W., Song, Y., Jin, Z., Zhao, H., Chen, C.: Unsupervised orthogonal facial representation extraction via image reconstruction with correlation minimization. Neurocomputing **337**, 203–217 (2019)
17. Sun, W., Zhao, H., Jin, Z.: A complementary facial representation extracting method based on deep learning. Neurocomputing **306**, 246–259 (2018)
18. Sun, W., Zhao, H., Jin, Z.: A visual attention based ROI detection method for facial expression recognition. Neurocomputing **296**, 12–22 (2018)
19. Tang, C., Yang, X., Lv, J., He, Z.: Zero-shot learning by mutual information estimation and maximization. Knowl. Based Syst. **194**, 105490 (2020)
20. Tschannen, M., Djolonga, J., Rubenstein, P.K., Gelly, S., Lucic, M.: On mutual information maximization for representation learning. arXiv preprint arXiv:1907.13625 (2019)
21. Wang, K., Peng, X., Yang, J., Lu, S., Qiao, Y.: Suppressing uncertainties for large-scale facial expression recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6897–6906 (2020)
22. Wen, L., Zhou, Y., He, L., Zhou, M., Xu, Z.: Mutual information gradient estimation for representation learning. arXiv preprint arXiv:2005.01123 (2020)
23. Xu, C., Dai, Y., Lin, R., Wang, S.: Deep clustering by maximizing mutual information in variational auto-encoder. Knowl. Based Syst. **205**, 106260 (2020)
24. Yang, H., Ciftci, U., Yin, L.: Facial expression recognition by de-expression residue learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2168–2177 (2018)