Computer Science and Information Systems 20(4):1869–1883 https://doi.org/10.2298/CSIS221228034J

Heterogenous-view Occluded Expression Data Recognition Based on Cycle-Consistent Adversarial Network and K-SVD Dictionary Learning Under Intelligent Cooperative Robot Environment

Yu Jiang[1] and Shoulin Yin[2,⋆]

[1] College of Fine Art and Design, Shenyang Normal University
Shenyang, 110034 China
zsshanln@163.com
[2] Software College, Shenyang Normal University
Shenyang, 110034 China
yslin@synu.edu.cn

Abstract. In space art design, the recognition of expression is of great help to the understanding of art. It is very difficult to obtain occlusion expression data from robot environment. In particular, it is very challenging to recognize the occluded expression. In the case of facial occlusion, it is difficult to extract the features of occluded expressions by traditional methods. In order to reduce the dependence of expression recognition on individuals, this paper proposes a cycle-consistent adversarial network and K-SVD dictionary learning method for occluded expression recognition in education management under robot environment. Firstly, the new method uses the cyclic-consistent generation adversarial network as the skeleton model, which can generate the un-occluded expression image without the need of paired data sets. Meanwhile, in order to improve the discriminant ability and image generation ability of the network, a multi-scale discriminator is used to construct the discriminant network. Then, the least squares and cyclic sensing loss are used to strengthen the constraints on the network model and improve the image quality. By subtracting the error matrix from the test sample, a clear image of the expression classification stage can be recovered. The clear image samples are decomposed into identity features and expression features by using the collaborative representation of two dictionaries. Finally, it is classified according to the contribution of each expression feature to the joint sparse representation. Experiments conducted on CK+, RAF-DB and SFEW datasets, the results show that the average accuracy of the new model is 98.44%, 87.12% and 62.17%, respectively. Compared with the traditional convolutional neural network models and advanced methods, this model effectively improves the accuracy of facial recognition in the case of facial occlusion.

Keywords: Heterogenous-view data, Intelligent Cooperative Robot, cyclic-consistent generation adversarial network, K-SVD dictionary learning.

1. Introduction

In recent years, Facial Expression Recognition (FER) has been widely used in human-computer interaction, automatic driving and mental health assessment. As a cross-domain

⋆ Corresponding author

technology, the development of facial expression recognition can promote the progress of face detection technology [1,2], face recombination technology [3], animation simulation technology [4] and other related technical fields. At the same time, it is also very important for the development of robotics, which can be applied in the robot face recognition, reduce costs and reduce the risk of personnel outstanding under the COVID-19 pandemic. Although existing facial expression recognition systems have a high recognition rate, most of them are obtained based on laboratory database systems, such as CK+, JAFFE, MMI, etc. Most of these face images are positive face images without any occlusion, so they are not universal in practical application. In order to improve the recognition rate of facial expressions in real scenes, researchers collected a large number of facial images to build an expression database [5-7], and proposed novel algorithms [8,9] and optimized network architecture [10]. However, according to the performance of the existing models in the database, the facial expression recognition technology in the real scene is still in its infancy. One of the biggest factors affecting the recognition rate is the occlusion problem. In the real scene, occlusion is inevitable. It may be caused by itself, such as self-occlusion caused by posture, hair, arms, etc., or by external objects, such as glasses, scarves, masks, food and other people's occlusion. This will inevitably lead to the decline of identification accuracy.

In the aspect of occluded facial expression recognition, researchers have proposed many methods to reduce the impact of occluded factors on facial expression recognition. The most commonly used methods include feature reconstruction, sub-region analysis, sparse representation, deep learning and so on. XBellamkonda et al. [11] jointly applied Robust PCA (RPCA) and significance detection to occlusion expression recognition. Firstly, significance detection was used to locate occlusions, and then RPCA was used to reconstruct the occlusive area of the image. Khan et al. [12] reconstructed 54 reference points in the occluded face image by combining the nearest point selection and fuzzy C-Means (FCM) algorithm. This feature recovery of the occluded area based on facial contour features relied heavily on reliable face detection and facial feature tracking, and required pre-positioning of the occluded area and accurate facial alignment, which was difficult to achieve in practical applications. Dapogny et al. [13] proposed a local expression prediction algorithm (LEPs) for expression classification prediction and AU prediction under the condition of local occlusion to reduce the negative impact brought by occlusion. Göre et al. [14] used the geometric features of face point displacement to predict the AU unit under oral occlusion. Gaussian mixture model (GMM) was used to model the gray pixel distribution of the face area to detect the occlusion area. Subarea analysis methods assume that the occlusion occurs in only a small part of the face. These methods usually produce satisfactory performance for small areas of occlusion. However, the granularity of subdividing a face into local areas and its effect on performance remains an open problem, especially for random occlusion without fixed location, shape, and size. At the same time, the method based on sub-region is sensitive to noise because of the inaccuracy of face location, alignment and normalization. Saka et al. [15] proposed a deep generative model, which used Markov Random Field (MRF) with gates as the first layer of DBN for expression recognition under occlusion. Houshmand et al. [16] proposed a deep neural network structure under the condition of partial occlusion, which used Gabor filters to extract multi-scale and multi-direction Gabor features from face images and input them into a three-layer Deep Boltzmann Machine (DBM) for emotion classification.

Yao et al. proposed that Wasserstein generative adversarial network could be used to complete the occlusion area of face image, which alleviated the impact caused by the loss of local expression information.

Deep learning techniques can automatically learn the most discriminative facial expression feature patterns from raw facial data, often without the need for a separate occlusion detection or reconstruction process [17-19]. However, using deep architectures requires large amounts of training data to ensure proper feature learning, the difficulty of adjusting large numbers of system parameters, and the need for expensive computations. Cen et al. [20] obtained various levels of occlusion base images by partitioning images, and constructed a non-orthogonal occlusion dictionary by using these images. Then the coefficients were obtained by sparse decomposition of the image to be measured and the expression category was determined in the subspace of the image to be measured. Wang et al. [21] proposed a robust regularized coding random occlusion expression recognition. By assigning different weights to each pixel of the expression image, the weight converged to the threshold through continuous generation selection. Finally, the sparse representation of the image to be measured was calculated by the optimal weight matrix. This method achieved good recognition effect. However, this method did not avoid the interference of identity characteristics on expression classification. The biggest advantage of sparse representation is that it is not only robust to occlusion and damage, but also can be used to estimate the occluded or damaged part of the face. The traditional sparse representation methods sparse representation (SRC) and collaborative representation (CRC) separate each dictionary atom and process it independently [22]. The relationships between atoms are not considered, and the resulting sparse representation is unstructured. Meanwhile, both SRC and CRC use the original training data set as the classification dictionary. When the training samples are damaged, blocked or deformed, the algorithm lacks robustness and the classification effect is poor. In low-rank algorithms, kernel norm is generally used to replace rank function, but kernel norm may not be able to effectively estimate the rank of matrix, and kernel norm is sensitive to Gaussian noise.

Our main contributions are as follows. This paper analyzes the occlusion problem in expression recognition and proposes a cycle-consistent adversarial network and K-SVD dictionary learning method with occlusion perception ability to extract facial expression features under robotic environment. The new method uses the cyclic-consistent generation adversarial network as the skeleton model, which can generate the un-occluded expression image without the need of paired data sets. Then, the least squares and cyclic sensing loss are used to strengthen the constraints on the network model and improve the image quality. By subtracting the error matrix from the test sample, a clear image of the expression classification stage can be recovered. The clear image samples are decomposed into identity features and expression features by using the collaborative representation of two dictionaries.

The structure of this paper is organized as follows. Section 2 introduces the related works. Section 3 detailed shows the facial expression method. Rich experiments are conducted in section 4. There is a conclusion in section 5.

2. Related Works

2.1. Generative Adversarial Network

Generative Adversarial Network (GAN) [23] is a deeply generative neural network model. The network model is composed of two functional modules, generator and discriminator. The structure of GAN is shown in Figure 1.

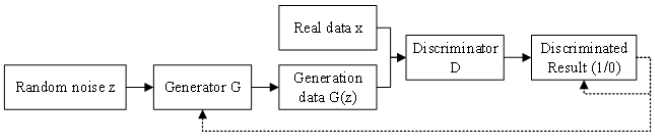

Fig. 1. Structure of GAN

The generator receives random noise z, gets the data distribution from the real sample and maps it to the new data space. It tries to generate a false sample $G(x)$ that can deceive the discriminator. The input of the discriminator consists of two parts, namely the data $G(x)$ generated by the generator and the real data x. The discriminator tries its best to determine whether the input sample is a generator generated sample or a real sample. If the test input is a real sample, the output is 1, otherwise it is 0. The generator and discriminator are constantly fighting and optimizing until the sample generated by the generator is fake and the discriminator has the lowest probability of discriminating wrong.

The training process of GAN can be regarded as a game between generator and discriminator. The training will first fix the generator G to update the parameters of the discriminator D. Then it fixes the discriminator D and updates the parameters of the generator G. The two models are iterated alternately and the optimal solution is finally reached. The antagonistic relationship between generator and discriminator is shown as follows:

$$\min_{G}\max_{D} V(D,G) = E_{x\sim p_{data}(x)}[lgD(x)] + E_{z\sim p_z(z)}[lg(1-D(G(z)))] \quad (1)$$

In order to minimize the error probability of discriminant model D, it is necessary to maximize the discriminant model when training the discriminant. When the input to the discriminator is the real sample x, $D(x)$ is expected to approach 1. When the input is a false sample $G(z)$, $D(G(z))$ is expected to approach 0, that is, $(1-D(G(z)))$ is expected to approach 1. For the generation model G, minimization is required, and the input of the generator is only random noise z. In this case, $D(G(z))$ is expected to approach 1, that is, $(1-D(G(z)))$ is expected to approach 0. The generated sample $G(z)$ is judged by the discriminator to be true with a probability value of 1. Only when $p_z = p_{data}$, the generator can learn the distribution of real sample p_{data}, and the discriminator's accuracy is stable at 0.5. At this time, the model can get the global optimal solution.

2.2. Sparse Representation Classification

Suppose n training samples can be divided into c categories, then the training set $D = [D_1, \cdots, D_c] \in R^{m \times n}$. In the sparse representation, test sample y can be approximately represented as a combination of feature vectors of training samples:

$$y = D_1 x_1 + D_2 x_2 + \cdots + D_c x_c \tag{2}$$

The sparse representation problem is to solve the coefficient vector under the constraint conditions of Equation (2) and minimize the L0 norm of x. According to the knowledge of compression theory, when Equation (2) is sparse enough, the minimization of L1 norm and L0 norm are equivalent, and the approximate solution of coefficient x can be obtained by solving the convex relaxation optimization problem in Equation (3).

$$\min_x ||x||_1, s.t. y = Dx \tag{3}$$

Sparse representation classification is more robust than traditional classifiers such as SVM, and abnormal noise points have less interference on classification.

3. Proposed Occluded Expression Recognition

Our proposed method is shown in figure 2. It mainly includes two parts: Cycle-Consistent Adversarial Network and K-SVD dictionary learning.

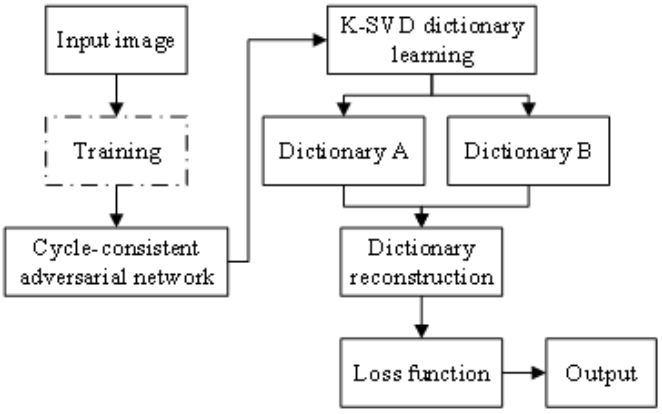

Fig. 2. Proposed occluded expression recognition framework

3.1. Cycle-Consistent Adversarial Network

CycleGAN (Cycle-Consistent Adversarial Network) is an unsupervised generative adversarial network based on the GAN network [24]. The network structure with two generators

and two discriminators is obtained by the mirror symmetry of traditional GAN. Because of this cyclic network structure, the CycleGAN model can convert images from two domains to each other without the need for paired image data sets. Its schematic diagram is shown in Figure 3. Where X and Y are the original domain X and target domain Y respectively. G and F are the generators of $X \rightarrow Y$ mapping and $Y \rightarrow X$ mapping respectively. D_X and D_Y are the corresponding discriminators. G converts the image in X domain to the image $G(x)$ in Y domain, and then the discriminator D_Y determines whether the image is true or false. Similarly, F converts the image in the X domain to the image $F(y)$ in the Y domain, and then the discriminator D_X determines whether the image is true or false.

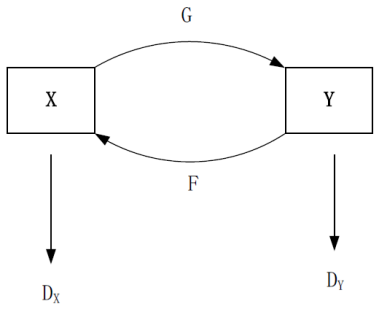

Fig. 3. Structure of CycleGAN

The generator G generates an image $G(x)$, and the discriminator D_Y determines if $G(x)$ is a real image. This process is a standard one-way GAN model training process. Therefore, the adversarial loss function of CycleGAN is consistent with that of GAN. The anti-loss function of G and D_Y is expressed as follows:

$$L_{GAN}(G, D_Y, X, Y) = E_{x \sim p_{data}(x)}[lgD_Y(y)] + E_{x \sim p_{data}(x)}[lg(1 - D_Y(G(x)))] \quad (4)$$

$$L_{GAN}(F, D_Y, Y, X) = E_{x \sim p_{data}(x)}[lgD_X(x)] + E_{y \sim p_{data}(y)}[lg(1 - D_X(F(y)))] \quad (5)$$

After the generator G converts the image x to $G(x)$ and then from F to $F(G(x))$, in order to ensure the consistency of the image after two networks as much as possible, that is, it tries to make $F(G(x)) \approx x$ and $F(F(y)) \approx y$, it needs to calculate the loss between the original image and $F(G(x))$. The cyclic uniform loss is calculated by the L1 norm of the original image and the mapped image, and its function expression is as follows:

$$L_{cyc}(G, F) = E_{x \sim p_{data}(x)}[||F(G(x)) - x||_1 + E_{y \sim p_{data}(y)}[||G(F(y)) - y||_1]] \quad (6)$$

Finally, the loss function of CycleGAN consists of the above three parts, and its total objective function is shown as follows:

$$L(G, F, D_X, D_Y) = L_{GAN}(G, D_Y, X, Y) + L_{GAN}(F, D_Y, Y, X) + \lambda L_{cyc}(G, F) \quad (7)$$

Where L_{GAN} is the counter loss. L_{cyc} is cyclic consistency loss. λ is the weight coefficient of cyclic consistency loss.

3.2. K-SVD Construction Based on Non-convex Logarithm Function Low-rank Decomposition

In expression recognition, there are similar features among the same expressions of different people, but there are also differences in identity, illumination, occlusion and so on. These details are not the key to determining categories. Adding these factors will reduce the classification accuracy of sparse representation. Therefore, the expression features and identity features of the expression images are separated by the non-convex logarithm function low-rank matrix decomposition. The dictionary learning is carried out respectively to obtain the double dictionary pattern of the in-class related dictionary and the differentially structured dictionary. The accuracy of sparse representation can be improved and the robustness of recognition can be enhanced by the decomposition of the internal structure information features of the image.

Given a training sample $X = [X_1, X_2, \cdots, X_c]$, it is assumed that it can be decomposed into a low-rank matrix L and a sparse matrix S. The objective function is as follows:

$$\min_{L,S}(rank(L) + \gamma||S||_{2,1}), s.t. X = L + S \quad (8)$$

In Equation (8), $rank(*)$ is the rank function. To prevent over-fitting, the constant $\gamma > 0$ is the weight factor used to balance the rank of the matrix and the sparsity of the noise matrix.

The rank function has non-convexity and discontinuity, and its minimization is an NP problem. Under certain conditions, the rank of matrix is replaced by the kernel norm convex approximation. The rank of the matrix is only related to the value of the largest non-zero singular value, whereas the kernel norm simply adds all the singular values of the matrix. When there are some large singular values, the kernel norm is too loose, so that the final calculation result will deviate from the real rank. Compared with the kernel norm, the non-convex approximation function can approximate the real rank of the matrix more accurately. In reference [25], when solving the subspace clustering problem, the logarithmic determinant function was used for non-convex approximation of the matrix rank function, and the experimental effect of cluster analysis in Yale B face Library was better than that of the kernel norm. It showed that using logarithmic determinant function to approximate the rank function of matrix was better, and had the advantages of strong anti-interference ability and small deviation.

The logarithmic approximation function is defined as formula (9).

$$||x||_d = logdet(I + (x^T x)^{0.5}) = \sum_{i=1}^{minm,n} log(1 + \sigma_i(x)) \quad (9)$$

In Equation (9), I is the identity matrix. σ_i is the singular value of matrix x. m and n are the dimensions of matrix x. The square root of $x^T x$ is taken to facilitate optimization.

Algorithm 1 Solving Lagrange function

Input: Training matrix X, parameter μ, γ, ε.
Output: L, S.

1: Fixed other values, update L.

$$L_{K+1} = \underset{L}{argmin} \sum_{i}^{minm,n} log(1+\sigma_i(L)) + \frac{\mu}{2}||\frac{Y_K}{\mu} + X - L - S^K||_F^2$$

2: Fixed other values, update S.

$$S_{K+1} = \underset{S}{argmin}\, \gamma||S||_{2,1} + \frac{\mu}{2}||\frac{Y_K}{\mu} + X - L - S^K||_F^2$$

3: Fixed other values, update Y.

$$Y_{K+1} = Y_K \gamma\mu(X - L_{K+1} - S_{K+1})$$

4: Check whether the convergence conditions satisfies:

$$||X - L_{K+1} - S_{K+1}||_F < \varepsilon$$

5: If it does not converge, the cycle continues until the convergence ends.

$$\underset{L,S}{min}\, ||L||_d + \gamma||S||_{2,1}, s.t. X = L + S \tag{10}$$

Formula (10) is solved by the augmented Lagrange multiplier method. The auxiliary variable Y is introduced for ease of calculation. Then the Lagrange function of Equation (10) is shown as follows:

$$L(L,S,Y,\mu) = \sum_{i=1}^{minm,n} log(\sigma_i(L)) + \gamma||S||_{2,1} + < Y, X - L - S > + \frac{\mu}{2}||X - L - S||_F^2 \tag{11}$$

In formula (11), Y is the Lagrange multiplier, and the initial value is 0. $< \cdot >$ represents the Euclidean inner product of a matrix. The specific solution steps are described as Algorithm 1.

Low-rank decomposition is carried out for each type of training sample. L and S are obtained after decomposition. L represents facial feature and S represents identity feature.

$$L = [L_1, L_2, \cdots, L_c] \in R^{n \times mc} \tag{12}$$

$$L = [S_1, S_2, \cdots, S_c] \in R^{n \times mc} \tag{13}$$

Where L represents the expression feature matrix. S stands for identity eigenmatrix. c is the number of emotion categories. Then the K-SVD dictionary learning algorithm is used to learn L and S to get dictionary A that can capture the discriminant features of the class and dictionary B that reflects the changes within the class. The K-SVD algorithm mainly improves the dictionary updating speed and reduces the complexity.

4. Experiments and Analysis

4.1. Datasets

In this paper, the experimental evaluation is conducted on three public databases, namely a laboratory image database CK+ [26] and two real scene databases, including RAF-DB [27] and SFEW [28].

The CK+ database is the most widely used database of laboratory collected images. CK+ consists of 593 video sequence items from 123 participants. These sequences range in duration from 10 to 60 frames and contain the transition of facial expressions from natural to peak. In the videos, 327 sequences from 118 gatherers are labeled with seven basic expressions based on the Facial Expression Action Coding System (FACS), including anger, contempt, disgust, fear, happiness, sadness and surprise. Because CK+ does not provide a specified training, validation, and test set, so the algorithm evaluation is not uniform on this database. Based on the static recognition method, the most commonly used data selection method is to extract the peak expression from the natural expression in the first frame to the last frame, and divide the subjects into n groups for n cross-validation experiments, in which n values are usually selected as 5, 8 and 10.

RAF-DB is the real world emotional face database, a large facial expression database, which has downloaded about 30,000 various facial images from the Internet. Each image is individually annotated about 40 times on the basis of crowd-sourcing. Subjects in the database vary greatly in age, gender, race, head posture, lighting conditions, occlusion (such as glasses, facial hair or self-occlusion), and post-processing operations (such as various filters and special effects). RAF-DB is characterized by diversity, large quantity and abundant annotation. The database contains two distinct subsets: the single-tag subset and the double-tag subset. The single label subset includes 7 types of basic emotions and boundary frames. The double-labeled subset includes 12 types of compound emotions, 5 accurate landmark locations, 37 automatic landmark locations, boundary boxes, races, age ranges, and each image gender attribute is annotated. Baseline classifiers are used to output basic emotions and compound emotions. In order to evaluate the performance of the test system objectively, the database is divided into training set and test set, in which the size of the training set is 5 times that of the test set, and the distribution of the two groups of expressions is almost the same.

SFEW database is a very challenging database of static frame pictures selected from actual expressions in real scenes. The SFEW database contains different levels of facial expression, unconstrained head postures, different shading, different age ranges, and different lighting variations. The sample database is an emotion tag with seven expressions: anger disgust, fear, happiness, sadness, surprise and nature. There are 95 topics in the database, of which there are 663 clearly labeled images. The database has been classified into training, validation, and test sets. SFEW 2.0 is divided into three groups: the training set with 958 samples, the verification set with 436 samples, and the test set with 372 samples. Each image in the database is classified into one of seven expressions, namely anger, disgust, fear, nature, happiness, sadness and surprise. The labels for the test and validation sets of the expression are public.

The decision unit training data set is the training occlusion decision unit. 100 facial images are selected from the open database for occlusion synthesis and used as training samples.

Occlusions are selected according to their frequency in daily life, such as fruit, hair, hats, books, cups, glasses, etc. In addition, another 100 natural occlusion images were selected as training samples. After consistent marking and inspection by team members, the database finally contains 200 training samples. In this paper, the threshold of area occlusion rate is set as 0.5, that is, when the area occlusion rate exceeds 1/2, the label is set as 0. As a result, the database is easy to tag and train. During the training process, the data expansion strategy is used to enhance the samples for 7 times, and the final accuracy rate reaches 85.4% through training.

4.2. Configuration of Experiment

The model proposed in this paper is based on Keras framework and runs on Ubuntu operating system. Experimental data obtained on NVIDIA CUDA framework 6.5 and NVIDIA GTX 1080GPU is used for experiments. In addition, GAN is used as the backbone network of proposed network. Firstly, the image data on ImageNet is used to initialize the network parameters. In the experiment, the small batch random gradient descent method is used to optimize the model. The initial value of the basic learning rate was set at 0.001, and the polynomial strategy was used to reduce it to 0.1. The momentum is set to 0.9 and the weight loss is set to 0.000 5. In the training phase, it sets the value of the actual batch size to 64 and selects the generation 10,000 times. When training the cyclic adversarial network, the face images on ImageNet are also used to pre-train the convolution layer, and all the parameters of the convolution layer are initialized. Then the parameters of the convolution layer are fixed and the fine-tuning training of the final fully connected layer is carried out. In the training process, the value of learning rate is set as 0.01. After 20,000 times of generation selection, the value of learning rate is adjusted to 0.0001 in the fine-tuning stage, and the iteration is continued for 10,000 times. The whole model training take 5 days, and after the parameters are fixed, the time for the model to process a single image is 1.2s.

4.3. Evaluation of Weight

This paper evaluates the weight factor α on three benchmark databases. During the test, the initial value of α is set to 0 and the increment was set to 0.1. When $\alpha = 0$, it means that only the output of the occlusion awareness network is used as the classification result. When $\alpha = 1$, it means that only the output of the generated adversarial network is used as the classification result. As shown in Table 1, the evaluation results are presented respectively on CK+, RAFDB and SFEW databases. In these three databases, when α values are 0.5, 0.7 and 0.6 respectively, the model achieves the best performance. Then, the value of α is further tested, and the results showed that when the value of α is 0.6, the overall performance of the model on the three databases is the best. Therefore, the value of α is finally set to 0.6 manually.

4.4. Experimental Results

The confusion matrix results of the proposed model on the three public databases are shown in Figure 4. Figure 4(a) shows the experimental results on the database CK+. For

Table 1. Evaluations on different datasets/%

α value	CK+	RAF-DB	SFEW
0	95.5	83.2	57.1
0.1	95.0	84.0	58.0
0.2	95.3	84.1	57.8
0.3	95.6	84.9	58.0
0.4	96.2	85.1	59.0
0.5	97.6	85.9	59.1

the 6 expressions, the recognition accuracy is above 95%, which is similar to the recognition rate of human beings. In particular, it is 99 percent accurate for the expressions with obvious changing features, such as happiness and anger. This means that the laboratory image database is no longer challenging for existing models. So researchers need to focus more on solving facial expression recognition problems in real situations. Figure 4(b) shows the confusion matrix in the database RAF-DB. For happy and angry expressions with obvious changing characteristics, the accuracy of the experimental model reaches more than 90%, which is within the acceptable range. For expressions that are not obvious in appearance, such as disgust and sadness, the recognition rate remains below 85%. The analysis of the experimental results shows that the main reason is that these expressions are not significantly different in the real scene. Even for humans, it is difficult to distinguish these two expressions accurately. Figure 4(c) shows the experimental results on the database SFEW, which is the most challenging database in the evaluation experiment. The experimental results show that only 2 emojis have a recognition rate of more than 80%, and all the other emojis have a recognition rate of less than 70%, or even only 49%. The analysis shows that the main reason for the low recognition rate of expression "disgust" is that the change of expression is not obvious; In addition, there are two pairs of expressions that are frequently mixed, including "sad" and "fearful," "surprised" and "happy." In reality these expressions are often associated with sex. For example, sadness caused by fear and surprise mixed with joy are expressions that are difficult for humans to distinguish from a single static image even in real situations.

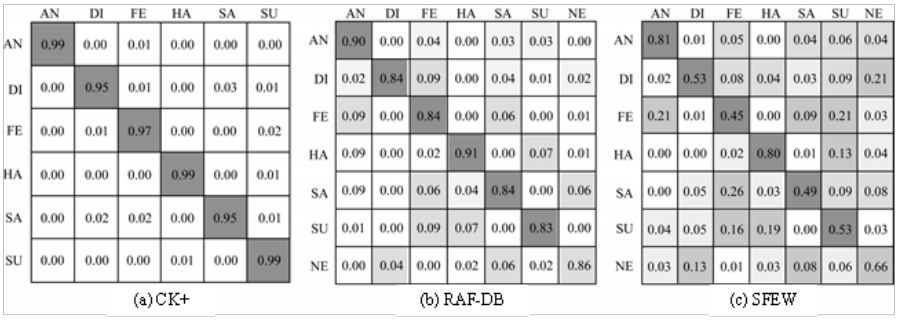

Fig. 4. Confusion matrices on three public databases

4.5. Comparison Results

In this paper, the experimental results of the proposed model are compared with those of similar models and widely used models including OFR [30], GMLA [31], DACS [32]. Tables 2 4 are the comparison results between the new model in this paper and other models in three different databases.

Table 2. Comparative evaluation of experiments on CK+ database

Model	Recognition rate/%	Time/s
OFR	92.15	12.5
GMLA	94.68	8.9
DACS	95.23	7.8
Proposed	97.44	3.5

Table 3. Comparative evaluation of experiments on RAF-DB database

Model	Recognition rate/%	Time/s
OFR	80.97	10.5
GMLA	82.94	9.1
DACS	84.64	5.7
Proposed	86.12	2.8

Table 4. Comparative evaluation of experiments on SFEW database

Model	Recognition rate/%	Time/s
OFR	55.41	12.7
GMLA	57.28	9.7
DACS	59.14	5.5
Proposed	62.17	3.8

Table 2 shows the comparison results on the database CK+. The analysis shows that the average accuracy of the model proposed in this paper is 97.33%. The images in CK+ database are all unshielded frontal face photos collected in the laboratory, and no specified training set and test set are provided. Therefore, different classification results will lead to different experimental results. It can also be seen from Table 2 that the recognition rate of most models in the laboratory database has achieved a satisfactory accuracy rate, achieving a recognition level comparable to that of human beings.

Table 3 shows the comparison results on RAF-DB database. The results show that the average accuracy of the proposed model is 86.12% and its accuracy is 0.93% higher than that of the OFR. Compared with GMLA, the model in this paper uses the sliding partition

selection method to partition in the drive domain, which does not depend on the accuracy of key points. The selection method is similar to the filter in the convolutional neural network, so that the network can retain more effective features. In addition, the single occlusion decision unit in this paper only needs to determine the occlusion threshold of the subarea, but does not need to calculate the precise occlusion proportion of the subarea, which makes the single occlusion decision unit easier to train. In the output fusion mode, the single factor mode is used for fusion, which further improves the performance of the model. In addition, compared with network B, it increases by 2.16%. The experimental results show that adding the zone occlusion decision unit and optimizing the network structure can improve the network performance.

Table 4 shows the comparison on the SFEW database. Through the analysis of the results of two real scene expression databases, RAFDB and SFEW, it can be seen that for the SFEW database with greater challenges, the performance improvement brought by the addition of occlusion decision unit is more obvious than that on the RAF-DB database, which confirms the practical application potential of occlusion awareness network. From the performance of the model in three open databases and the comparison results with the existing methods, the model has practical application value. Also the time consumption is the lowest with proposed method.

5. Conclusion

In this paper, we propose a cycle-consistent adversarial network and K-SVD dictionary learning method for occluded expression recognition in education management under robot environment. This method uses the low-rank decomposition of non-convex logarithm approximate rank function to separate the identity features from the facial features, and then carries on the dictionary learning to get a double dictionary. The identity features and expression features of the test images are reconstructed by using double dictionaries. And it introduces the cyclic generative adversarial network as the basic network model. The multi-scale discriminator is used to replace a single discriminator to improve the quality of the generated images, so that the generated images are more natural and closer to the real face images, so that the impact caused by local occlusion can be reduced during recognition. Finally, the classification is realized according to the contribution of each facial feature in the joint sparse representation. The experimental data in CK+ and KDEF show that the algorithm in this paper can overcome the negative effects brought by individual differences to a certain extent, which not only improves the robustness of sparse representation, but also has strong robustness to face recognition with different regions and different degrees of occlusion. The next step will be to process and recognize multi-pose facial images.

Acknowledgments. This work was supported by ”Horizontal project: Appearance design of intelligent cooperative robot SCR series”. The authors gratefully acknowledge the anonymous reviewers for their valuable comments.

References

1. Kumar A, Kaur A, Kumar M. Face detection techniques: a review[J]. Artificial Intelligence Review, 2019, 52(2): 927-948.

2. Jiang M, Yin S. Facial expression recognition based on convolutional block attention module and multi-feature fusion[J]. International Journal of Computational Vision and Robotics, 2023, 13(1): 21-37.
3. Chi C, Zhang S, Xing J, et al. Selective refinement network for high performance face detection[C]//Proceedings of the AAAI conference on artificial intelligence. 2019, 33(01): 8231-8238.
4. Li X, Lai S, Qian X. Dbcface: Towards pure convolutional neural network face detection[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 32(4): 1792-1804.
5. Zhang S, Wen L, Shi H, et al. Single-shot scale-aware network for real-time face detection[J]. International Journal of Computer Vision, 2019, 127(6): 537-559.
6. Jisi A and Shoulin Yin. A New Feature Fusion Network for Student Behavior Recognition in Education[J]. Journal of Applied Science and Engineering. vol. 24, no. 2, pp.133-140, 2021.
7. Tang X, Du D K, He Z, et al. Pyramidbox: A context-assisted single shot face detector[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 797-813.
8. Zhang S, Wang X, Lei Z, et al. Faceboxes: A CPU real-time and accurate unconstrained face detector[J]. Neurocomputing, 2019, 364: 297-309.
9. Wang X. Crowd Density Estimation Based On Multi-scale Information Fusion And Matching Network In Scenic Spots[J]. Journal of Applied Science and Engineering, 2022, 26(6): 865-875.
10. Singh P, Chaudhury S, Panigrahi B K. Hybrid MPSO-CNN: Multi-level particle swarm optimized hyperparameters of convolutional neural network[J]. Swarm and Evolutionary Computation, 2021, 63: 100863.
11. Bellamkonda S, Gopalan N P, Mala C, et al. Facial expression recognition on partially occluded faces using component based ensemble stacked CNN[J]. Cognitive Neurodynamics, 2022: 1-24.
12. Khan M A, Arshad H, Nisar W, et al. An integrated design of fuzzy C-means and NCA-based multi-properties feature reduction for brain tumor recognition[M]//Signal and image processing techniques for the development of intelligent healthcare systems. Springer, Singapore, 2021: 1-28.
13. Dapogny A, Bailly K, Dubuisson S. Confidence-weighted local expression predictions for occlusion handling in expression recognition and action unit detection[J]. International Journal of Computer Vision, 2018, 126(2): 255-271.
14. Göre E, Evlioğlu G. Assessment of the effect of two occlusal concepts for implant-supported fixed prostheses by finite element analysis in patients with bruxism[J]. Journal of Oral Implantology, 2014, 40(1): 68-75.
15. Saka K, Kakuzaki T, Metsugi S, et al. Antibody design using LSTM based deep generative model from phage display library for affinity maturation[J]. Scientific reports, 2021, 11(1): 1-13.
16. Houshmand B, Khan N M. Facial expression recognition under partial occlusion from virtual reality headsets based on transfer learning[C]//2020 IEEE Sixth International Conference on Multimedia Big Data (BigMM). IEEE, 2020: 70-75.
17. Shi Q, Yin S, Wang K, et al. Multichannel convolutional neural network-based fuzzy active contour model for medical image segmentation[J]. Evolving Systems, 2022, 13(4): 535-549.
18. Shen W. A Novel Conditional Generative Adversarial Network Based On Graph Attention Network For Moving Image Denoising[J]. Journal of Applied Science and Engineering, 2022, 26(6): 831-841.
19. Gao S. A two-channel attention mechanism-based MobileNetV2 and bidirectional long short memory network for multi-modal dimension dance emotion recognition[J]. Journal of Applied Science and Engineering, 2022, 26(4): 455-464.
20. Cen F, Zhao X, Li W, et al. Deep feature augmentation for occluded image classification[J]. Pattern Recognition, 2021, 111: 107737.
21. Wang K, Peng X, Yang J, et al. Region attention networks for pose and occlusion robust facial expression recognition[J]. IEEE Transactions on Image Processing, 2020, 29: 4057-4069.
22. Rubinstein R, Bruckstein A M, Elad M. Dictionaries for sparse representation modeling[J]. Proceedings of the IEEE, 2010, 98(6): 1045-1057.

23. Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139-144.
24. Wen L, Wang Y, Li X. A new Cycle-consistent adversarial networks with attention mechanism for surface defect classification with small samples[J]. IEEE Transactions on Industrial Informatics, 2022, 18(12): 8988-8998.
25. Peng C, Kang Z, Li H, et al. Subspace clustering using log-determinant rank approximation[C]//Proceedings of the 21th ACM SIGKDD international conference on Knowledge Discovery and Data Mining. 2015: 925-934.
26. Chen W. A novel long short-term memory network model for multimodal music emotion analysis in affective computing[J]. Journal of Applied Science and Engineering, 2022, 26(3): 367-376.
27. Li S, Deng W, Du J P. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2852-2861.
28. Wang K, Peng X, Yang J, et al. Region attention networks for pose and occlusion robust facial expression recognition[J]. IEEE Transactions on Image Processing, 2020, 29: 4057-4069.
29. Poux D, Allaert B, Ihaddadene N, et al. Dynamic facial expression recognition under partial occlusion with optical flow reconstruction[J]. IEEE Transactions on Image Processing, 2021, 31: 446-457.
30. Zhao Z, Liu Q, Wang S. Learning deep global multi-scale and local attention features for facial expression recognition in the wild[J]. IEEE Transactions on Image Processing, 2021, 30: 6544-6556.
31. Farzaneh A H, Qi X. Facial expression recognition in the wild via deep attentive center loss[C]//Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2021: 2402-2411.

Yu Jiang, female, associate professor, postgraduate student, visiting scholar in Sheffield Hallam University in April 2012, CEA graphic Designer, (double professional) National Vocational Skills Appraisal of decorative art assessment. Researchr direction: Environmental art design.

Shoulin Yin is with Software College, Shenyang Normal University, Shenyang 110034, China. Research direction: image processing and AI.

Received: December 28, 2022; Accepted: June 04, 2023.