Hindawi
Complexity
Volume 2020, Article ID 8855885, 10 pages
https://doi.org/10.1155/2020/8855885

Research Article

Feature Guided CNN for Baby's Facial Expression Recognition

Qing Lin,[1] **Ruili He**,[2] **and Peihe Jiang**[3]

[1]*Integrated Information Center of Yantai, Yantai 264003, China*
[2]*School of Computer Science and Technology, Harbin Institute of Technology, Weihai 264209, China*
[3]*School of Opto-Electronic Information Science and Technology, Yantai University, Yantai 264005, China*

Correspondence should be addressed to Peihe Jiang; jiangpeihe@163.com

Received 5 September 2020; Revised 23 October 2020; Accepted 5 November 2020; Published 23 November 2020

Academic Editor: Min Xia

State-of-the-art facial expression methods outperform human beings, especially, thanks to the success of convolutional neural networks (CNNs). However, most of the existing works focus mainly on analyzing an adult's face and ignore the important problems: how can we recognize facial expression from a baby's face image and how difficult is it? In this paper, we first introduce a new face image database, named BabyExp, which contains 12,000 images from babies younger than two years old, and each image is with one of three facial expressions (i.e., happy, sad, and normal). To the best of our knowledge, the proposed dataset is the first baby face dataset for analyzing a baby's face image, which is complementary to the existing adult face datasets and can shed some light on exploring baby face analysis. We also propose a feature guided CNN method with a new loss function, called distance loss, to optimize interclass distance. In order to facilitate further research, we provide the benchmark of expression recognition on the BabyExp dataset. Experimental results show that the proposed network achieves the recognition accuracy of 87.90% on BabyExp.

1. Introduction

Facial expressions play an important role in human being's communication. The ability to differentiate genuine displays of emotional experience from the posed ones is very important for dealing with day-to-day social interactions. Humans and computer algorithms can greatly benefit from being able to distinguish the genuine expression from the posed one. Possible applications of automated facial expression recognition include better transcription of videos, movies, or advertisement recommendations and detection of pain in telemedicine. Therefore, facial expression recognition has attracted a vast amount of attention in the past two decades [1–6]. The development of facial expression recognition relies heavily on an adequate database of facial expressions. However, due to the nature of facial expressions, there are a limited number of publicly available databases providing a sufficient number of facial images tagged with accurate expression information. Table 1 shows the major differences of the existing image databases with the number of images, number of subjects, expression distribution, data size, and the released years. However, most of the existing works and datasets [7–11] focus on analyzing adult faces, which ignore how to analyze facial expressions from baby facial images. Although some datasets include children, there are actually very few images of very young children. None of these datasets is specifically designed to explore the expression of babies. There are two main reasons for the lack of research on baby face analysis. The first reason is that the community has not realized the application values of analyzing baby's facial expression. In fact, there are many applications of analyzing the facial expressions of babies, such as advertising marketing for parents, intelligent family child care, and scientific parenting. The second reason may be traced to the additional challenge of obtaining the baby face datasets with accurate expression labels.

As we all know, 0–2 years old is a golden period for the development of a baby and for laying a solid foundation for their lifelong physical and mental health. Therefore, it is valuable to develop the algorithm to interpret a baby's facial expressive signals for scientific parenting. In addition, due to the support of national policies and people's growing

Table 1: Overview of the existing facial expression datasets.

Datasets	JAFFE	PIE	MMI	BU-3DFE	CK+	FER	CAFE	SFEW	RAF-DB
Images	213	40,000	740	2500	593	35,887	1192	1766	29,672
Subjects	10	68	25	100	137	—	100	—	—
Class	7	4	—	7	7	7	7	7	7
Size	256 x 256	—	720 x 576	—	640 x 480	48 x 48	Square	720 x 576	—
Age	—	—	19–62	18–70	18–50	—	2–8	—	0–70
Gender	Female	Both	Both	Both	Both	Both	Both	Both	Both
Year	1988	2000	2005	2006	2010	2013	2014	2015	2017

attention to the growth and development of a baby, the parenting market has been expanding. Accurate recognition of facial expressions of a baby is of great significance to facilitate the development of scientific parenting. All these real needs have brought a strong motivation to the study of recognizing baby's face expressions.

Recently, researchers have realized the importance of children's facial expressions in order to study developmentally the interpretation of these expression datasets. For example, the new NIMH Children's Emotional Face Picture Collection (NIMH-ChEFS) contains photos of children aged 10–17 [12], the Radboud Faces Database includes photos of 8- to 12-year-olds [13], and the CAFE set features photographs of 2- to 8-year-old children [14]. Although these new datasets give researchers the option to use a sample of children aged 2–17 years, there have been no datasets that feature smaller children to date. On the contrary, all the datasets mentioned above for children's facial expressions have only a small number of images, which are not suitable for training convolutional neural network (CNN) models. In addition, these datasets contain the facial images with posed expressions in a lab-controlled environment.

In this paper, to address the aforementioned issues, we propose a new image dataset with expression labels of baby faces for automatic facial expression recognition. Our dataset, which is called the BabyExp dataset, contains more than 12,000 images from babies younger than two years old showing spontaneous expressions in an uncontrolled environment. Each face image is annotated with one of three facial expressions (i.e., happy, sad, and normal). It is complementary to existing adult face datasets and can shed some light on exploring baby face analysis. Our key contributions are summarized as follows:

(1) We present a facial expression dataset, named BabyExp, which contains more than 12,000 images from babies showing spontaneous genuine expressions in an uncontrolled environment. Each image is annotated with one of three facial expressions (i.e., happy, sad, and normal).

(2) We propose a new distance loss function to effectively enhance the discriminative ability of distance between classes in unconstrained facial expression recognition tasks.

(3) In order to facilitate further research, we proposed a new method for facial analysis and evaluated its performance on the BabyExp dataset. Experimental results show that the proposed network achieves a recognition accuracy of 87.90% on the test set of BabyExp.

2. Materials and Methods

2.1. Data Collection. Our baby face images are generated from both static images and video sequences uploaded by parents using smartphones. We will introduce the preprocessing of the BabyExp dataset in the following. For the original images and the original video data, we first perform face detection, then perform face cropping, and finally perform picture similarity detection. A detailed description can be found in the following.

2.1.1. Image Preprocessing. For image processing, we first use the Dlib visual library [15] and the OpenCV visual library to perform face detection and cropping on the original image. During the face detection, we adopt the following strategy. First of all, if a face appears, the face section will be extracted. Second, if no face is detected during the detection, we rotate the image 270 degrees clockwise at 90 degrees each time. If a face appears during the three rotation detection processes, then we crop and save the face image. Last, if there are two or more faces detected in the image, we will assume that this image will have an adult face or a face that is not a human face but is misidentified as a human face. Then, we will discard such images.

It is important to note that the area of the original picture of the baby's face is not very large. At this point, the picture is redundant. If it is used directly for training, the model converges slowly, resulting in poor test results. In order to reduce the large amount of nonface information in the image, therefore, after using the above Dlib face detection strategy, when cropping the face, we crop the face area according to a specific artificial strategy and save it. The main purpose is to obtain a noise-free and good-quality baby face image dataset in order to obtain a better model during the training process and a better accuracy during the test process. We then crop the original image according to the new picture size and finally normalize the cropped image (the normalized size is 256×256).

2.1.2. Video Preprocessing. We segment the original video data, take an image every 30 frames, and then perform the same process as the static image data preprocessing on the images from the video frames, detecting, rotating, and finally cropping the baby's face picture. It should be noted that

because the pictures obtained by intercepting video frames may have great similarities, many images are redundant, so the only different operation different from the static image is that, after the picture is cropped and saved, we need to perform picture similarity matching operations to filter the image. We use SSIM [16] to perform similarity matching and specify to delete images with similarity greater than 90%.

2.2. Data Annotation. After preprocessing, we get 7,600 images, and we will tag the images with facial expressions. Because babies are all at the stage of 0–2 years old, their expressions are not as diverse as those of the adults. For this reason, we specially selected three main baby expressions (i.e., normal, sad, and happy) for the BabyExp dataset. The marking process is divided into three steps: manual labeling, label statistical analysis, and label aggregation.

In the manual labeling step, 10 raters coming from Harbin Institute of Technology were selected to manually label the data. Without given any information, the subjects were asked to classify the photos according to their own experience. In order to save time and to boost classification efficiency, we used C++ language to design a manual labeling tool for manual classification and record the human evaluator choice of the expression label. For each input image, we asked 10 raters to label the image into one of 3 emotion types and 1 error fold: happy, sad, normal, and error. The raters are required to choose one single emotion for each image. After labeling, there will be four categories, i.e., happy, normal, sad, and error. The error category represents that an image is not a human face or the face is unclear.

The second step is to label statistical analysis. After the manual labeling of 10 people is completed, it is necessary to analyze the expressions in all the categories. The statistical result is an expression category selected by 10 people per picture. With labels from 10 raters for each face image, we can generate a probability distribution of emotion captured by the facial expression. Let N denote the number of the training examples $I_i, i = 1, \ldots, N$. Given the i-th example I_i, its label distribution from the raters can be expressed as $p_k^i, k = 1, \ldots, 4$. Naturally, we have

$$\sum_{k=1}^{4} p_k^i = 1. \tag{1}$$

The final step is to aggregate the labels of each image. After the second step, we need to aggregate the label of each expression generated by the 10 people. The combined labeling results are happy, normal, sad, and error. In most of the existing facial expression datasets, each facial image is only associated with one single label. If the image has more than one label, it is natural to assign the image to the label of the largest p_k^i. We experimented majority voting schemes. More formally, we create a new target distribution.

$$\widehat{p}_k^i = \begin{matrix} 1, & \text{if } k = \operatorname{argmax}_j p_j^i, \\ 0, & \text{otherwise.} \end{matrix} \tag{2}$$

After processing, when encountering an image, a certain type of expression will be selected, which means that the image is the corresponding category. If an image has the same labeling number of people and both have the maximum number of votes, the image is not classified, and they are marked twice to determine the baby's expression label of the image. Finally, in the end, we obtained 2,502 happy images, 4,028 normal images, and 1,070 sad images, as shown in Figure 1. It can be clearly seen that the three expression distributions in the baby expression dataset are unbalanced. This is because babies are different from adults who have rich expressions leading to a uniform expression distribution. Since expressions of babies from 0 to 2 years old are still developing and the expression types are relatively monotonous, especially in the absence of outside interference, most of the time, the baby is in a calm state followed by the state of laughter and finally, the state of sadness, so we can see that the proportion of normal is relatively large, and the proportion of sad is relatively small, which is very consistent with the expression characteristics of the baby, but imbalanced data may have a strong impact on the accuracy of the research experiment results; one solution is to use data augmentation and synthesis to balance the distribution of classes during the preprocessing phase.

2.3. Data Augmentation. According to the dataset information obtained above, there is an imbalance in the dataset, which will adversely affect the subsequent experimental work. Although deep learning has a strong characteristic learning ability, some technical hurdles prevent their successful applications to our dataset. First, deep neural networks require a lot of training data to avoid overfitting. Additionally, models trained using imbalance facial expression samples have a poor generalization ability and are prone to overfitting, which is illustrated in the experiments we introduced later in the experimental section. So, we need to perform data augmentation to promote data balance and facilitate the use of deep learning methods for experiments.

At present, generative adversarial networks (GANs) [17] are a popular research method in the field of machine learning. Their basic idea is derived from the game of two players in game theory. In the GAN framework, a "generator" network is tasked with fooling a "discriminator" network into believing that its own samples are real data. Inspired by the successful application of the GAN in the field of image style transfer, this project will use the GAN as a network model for image enhancement processing. We can use the resulting generative model to generate faces with specific expressions from nothing but random noise. Many different types of GANs require paired datasets for image style transfer. Baby expression images do not have paired data for sad and happy expressions corresponding to the same normal expressions of the baby, so the research contents in this part will draw on the important idea of CycleGAN [18] asymmetry training for unpaired image-to-image translation. The research contents in this part mainly include data augmentation of sad and happy facial expression images for imbalanced baby facial expression data based on CycleGAN.

The CycleGAN architecture contains two generators and two adversarial discriminators: Generator A, Generator B,

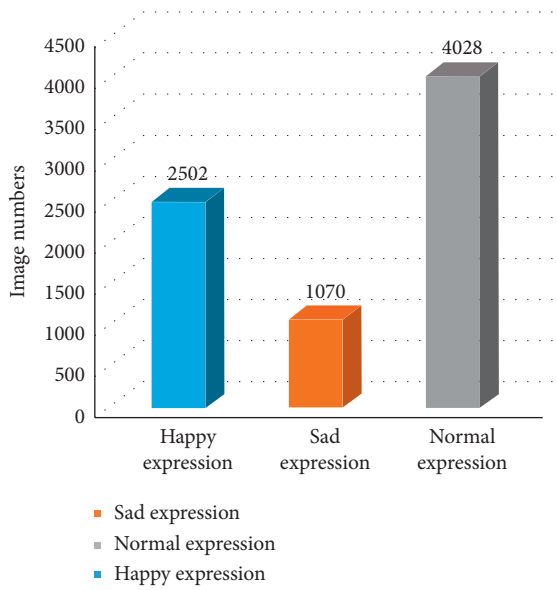

FIGURE 1: Image numbers of three expressions after preprocessing.

Discriminator A, and Discriminator B, where Generator A tries to generate images Generated_B that look similar to images from domain B, while Discriminator B aims to distinguish between translated samples Generated_B and real samples B. The overall structure of the algorithm in our data augmentation design is shown in Figure 2. Generator A inputs normal expression image A and output happy expression image Generated_B. Cyclic_A generated by Generator B brings Generated_B back to the original normal expression image A, where Cyclic_A is called the cyclic image of A. Generator B inputs happy expression image B and outputs normal expression image Generated_A. Cyclic_B is generated by Generator A, and Generated_A is brought back to the original happy expression image B. Cyclic_B is called the circular image of B. Discriminator A is used to distinguish true or false of the input normal expression image, and Discriminator B is used to distinguish true or false of the input happy expression image, respectively. Similarly, the data augmentation of sad expressions has the same process structure as that of happy expressions, which is not described in detail here.

It must be pointed out that because the number of normal expressions is sufficient, we have only enhanced the sad and happy expression image data. Finally, after data augmentation of CycleGAN, 1,498 happy expression images and 2,955 sad expression images are finally selected and generated. The total amount of facial expression data we obtained is shown in Table 2. It can be seen that, after data augmentation, we obtained 4,000 happy images, 4,028 normal images, and 4,025 sad images. We have a total of 12,053 baby facial expression images. We call it the BabyExp dataset, of which 4,453 are generated images. The amount of data for three facial expressions has reached an equilibrium state for the future academic research.

2.4. Proposed Methods. The overall pipeline of the proposed deep learning approach is depicted in Figure 3. Our proposed framework, called VFESO-DLSE, is composed of four modules: feature extraction, feature refinement, covariance pooling, and CNN classification. We also propose a new loss function, called distance loss, denoted as $\mathscr{L}_{DL}$.

2.4.1. Distance Loss. Min Xia et al. [19] found that the feature constraint helps enlarge the feature distance of different age range feature space in face images with similar feature distributions. Inspired by this, we propose a novel loss function, called distance loss, which takes strong feature constraint into baby facial expression learning. The distance loss aims to learn representations with lower intraclass variations and higher interclass distances. As we all know, by pushing the samples to the corresponding class center in the feature space during the training, the center loss [20] significantly reduces the intraclass difference. The center loss is defined as the sum of the square distance between the sample and its corresponding class center in the feature space. The center loss is denoted as $\mathscr{L}_C$:

$$\mathscr{L}_C = \frac{1}{2}\sum_{i=1}^{m}\left\|x_i - c_{y_i}\right\|^2, \tag{3}$$

where y_i is the class label of the i-th sample; x_i denotes the feature vector of the i-th sample taken from the FC layer before the decision layer; c_{y_i} denotes the center of all the samples with the same class label as x_i; and m is the number of samples in the mini-batch. Our distance loss denoted as $\mathscr{L}_{DL}$ is defined as

$$L_{DL} = L_C + \lambda_1 \sum_{C_j \in N_j} \sum_{\substack{C_k \in N_k \\ C_k \neq C_j}} \left(\frac{1}{\left\|C_k - C_j\right\|_2 + 1}\right), \tag{4}$$

where N_j and N_k denote the set of expression labels and C_k and C_j denote the k-th and j-th centers. Specifically, the first term was used to narrow the distance between the sample and the center of the corresponding class, and the second term was used to punish the similarity between different expressions. λ_1 is used to balance the weights of the two terms. By minimizing the distance loss function, the same expression will be brought closer, and different expressions will be pushed in the feature space.

2.4.2. Feature Guided CNN. As we all know, the expression change of babies aged 0 to 2 years will be less distorted. Although CNNs have achieved great performance in image processing [21–23], traditional CNNs consist of fully connected layers, maximum or average poolings, and convolutional layers to capture only first-order information [24]. We believe that second-order statistics is more suitable to capture such baby's expression distortions than first-order statistics. So, we take network architecture model-4 presented in [25] as a baseline model. Related studies [26, 27] have proved that the trained deep convolutional network can be used as a feature extraction tool for classification tasks, and it has a generalization ability. Following up this idea, we apply the famous VGG16 [28] model for feature extraction

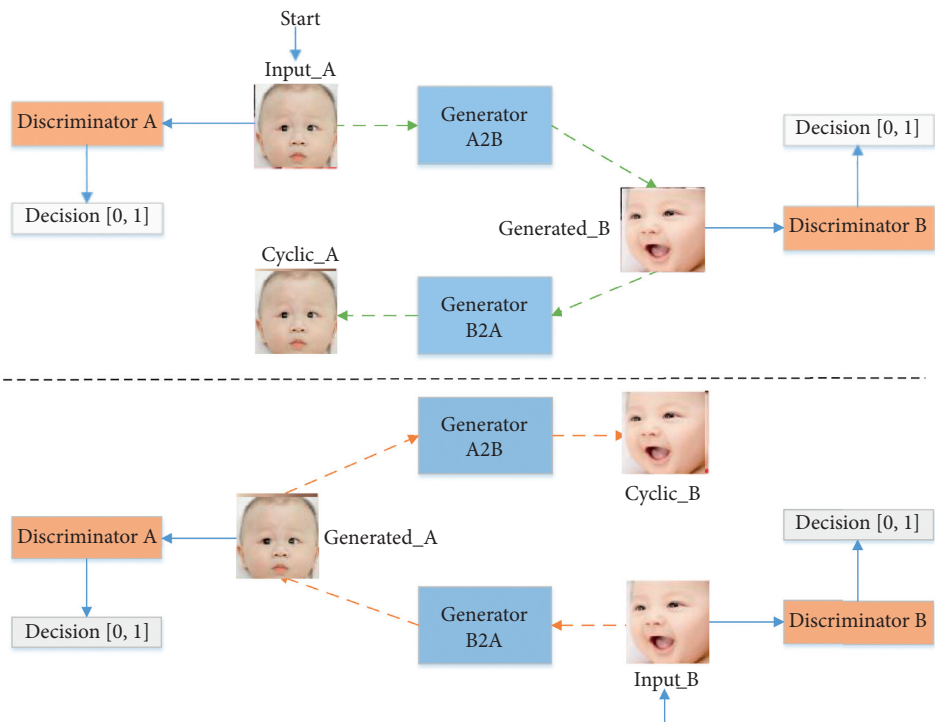

Figure 2: Example of happy expression image data augmentation process.

Table 2: Expression details for the BabyExp dataset.

Total images	Expression class	Age (month)	Happy images	Sad images	Normal images
12,053	3	0–24	4000	4025	4028

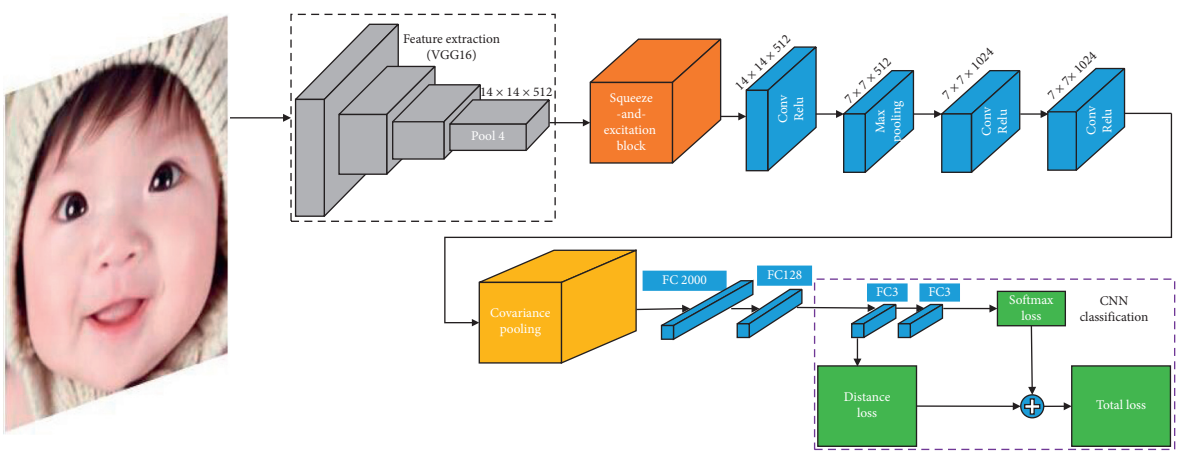

Figure 3: The overall architecture of the proposed deep learning approach VFESO-DLSE.

in our method. VGG16 is a typical CNN model. It has 13 convolutional layers, 5 pooling layers, and 3 fully connected layers for face recognition. To extract expression features, we use a pretrained VGG16 network on the expression dataset to extract features (referred to as VFE). For each facial image, we use the $14 \times 14 \times 512$ size feature maps of the fourth pooling layer to represent an image feature.

For the feature refinement stage, we use the squeeze-and-excitation (SE) block [29] to refine the CNN functionality and highlight the regions of expression that need to be highlighted, thereby explicitly modeling the interdependencies between the channels by adaptively recaliberating the channel's feature response. The detailed structure can be seen in Figure 4, and γ is a scaling parameter (16 in this paper). The purpose of this parameter is to reduce the number of channels and thus reduce the computation. C represents the number of channels, and H, W represent the height and width of the feature map

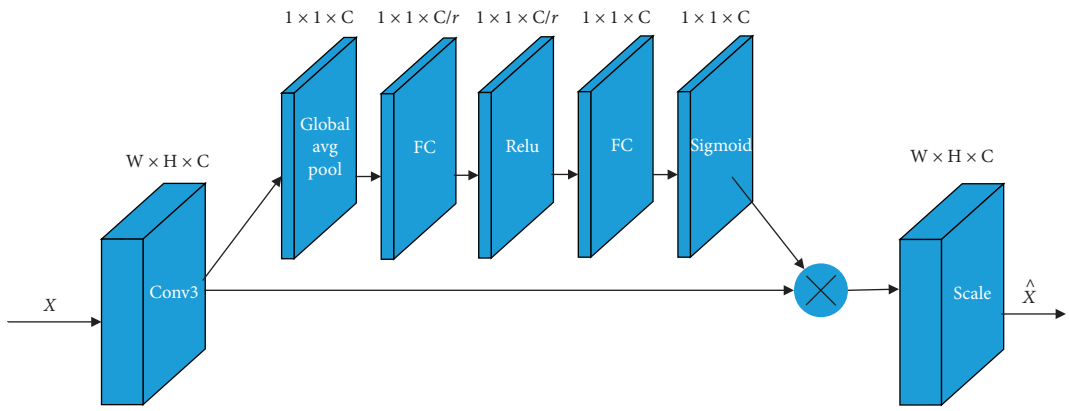

Figure 4: Squeeze-and-excitation (SE) block; γ is a scaling parameter.

input from the previous layer. The SE module first performs a squeeze operation on the feature map obtained by the convolution to obtain channel-level global features; here, we use global average pooling as the squeeze operation. Then, an excitation operation is performed on the global features. Two fully connected layers form a bottleneck structure, and the correlation between the channels is modeled. The number of output weights is the same as the number of input features. As shown in Figure 4, we first reduce the feature dimension to 1/16 of the input and then activate it through ReLu and then rise back to the original dimension through a fully connected layer, which learns the relationship between each channel and also obtains the weight of different channels and finally multiplies the original feature map to get the final feature. In essence, the SE module performs attention or gating operations on the channel dimension. This attention mechanism allows the model to pay more concern about the channel features with the most information and suppress those unimportant channel features.

Then, three convolutions with kernel size 3×3 are followed, and we use ReLU [20] as the activation function for each convolution layer and two max pooling layers. Then, the same as baseline [25], we also use covariance pooling after the last convolutional layer and before the fully connected layers. In the last classification part, the total loss of our network architecture training is formulated as follows:

$$\mathscr{L}_{\text{total}} = \mathscr{L}_s + \lambda \mathscr{L}_{DL}, \tag{5}$$

where $\mathscr{L}_s$ denotes the softmax loss and $\mathscr{L}_{DL}$ denotes the distance loss. The hyper parameter λ is used to balance the two loss functions.

2.5. Experiments

2.5.1. Experimental Setup. All the training and testing are carried out on the NVIDIA GeForce GTX 1080Ti 16G GPUs. We use deep learning framework TensorFlow [30] to develop the model. On an Ubuntu Linux system with NVIDIA GPUs, it takes 10–15 hours to train a model based on our network structures.

2.5.2. Implementation Details. We set up three major experiments: the first experiment is to evaluate the state-of-the-art adult facial expression analysis methods on BabyExp to see if the adult expression recognition method works for baby images. In this part, we use the methods trained on SFEW2.0 and test on BabyExp, and Table 3 shows the results of this experiment.

The second experiment is to demonstrate the effectiveness of the proposed method VFESO-DLSE. We compare our method against four designed architectures: DLP [31], the baseline [25], baseline + distance loss (SO-DL), and baseline + distance loss + SE block (SO-DLSE) (the structure can be seen in Figure 5). It should be noted that since our baseline network is based on the model from [31], we trained and tested the experimental results from scratch with our own BabyExp dataset for better comparison. Same as in [25], here, we use the center loss [32] in any case to train the network, not the locality preserving loss [31], because we do not deal with compound emotions. Table 4 shows the results of this experiment. In order to objectively measure the performance, the BabyExp dataset is divided into training and test sets, where the test set contains 2,413 images, and the remaining 9,640 images are used as the training set. The dataset is then resized to a fixed size 100×100, which is subsequently sent to the CNN classifier for expression recognition. It should be noted that the image size is resized to 224×224 only when entering the VFESO-DLSE method. The labeled facial expression dataset is quite small; thus, we use the conventional data augmentation method to generate more training data. In the data augmentation stage, we augment the set of training images in BabyExp by random flipping, rotating each with $\pm 10^{\circ}$, and random crop. We then train our networks for 700 epochs with the following parameters: learning rate 0.0001–0.005, weight decay 0.05, momentum 0.9, batch size 128, and linear learning rate decay in the

Table 3: Experimental results of adult expression recognition models on the adult and BabyExp dataset.

Models	SFEW2.0	BabyExp
DLP (trained on SFEW2.0)	54.45	39.70
Baseline (trained on SFEW2.0)	58.14	40.78

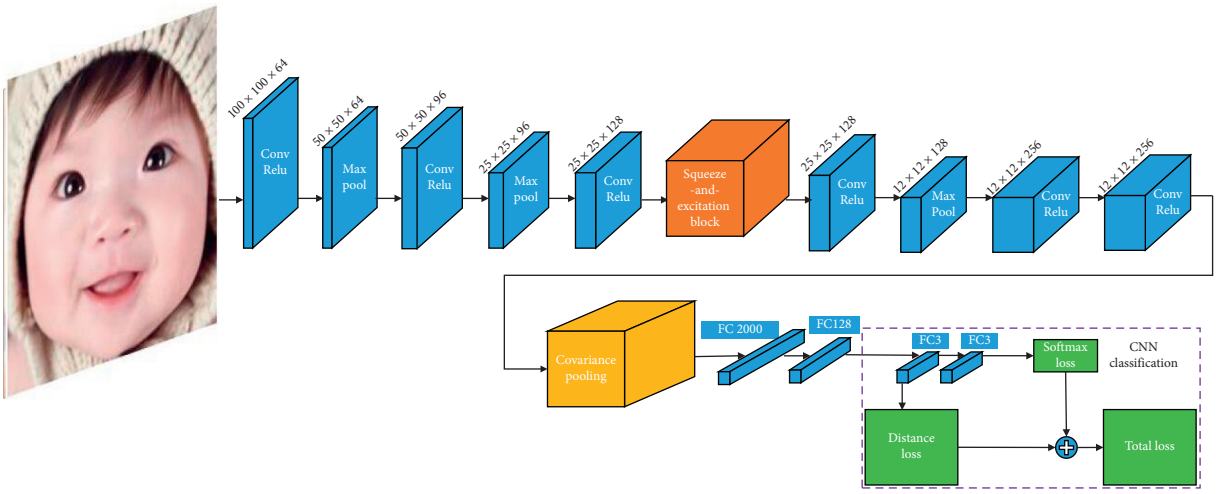

Figure 5: The overall architecture of the deep learning approach SO-DLSE.

Table 4: The expression recognition performance of different methods on the BabyExp dataset (trained from scratch).

Models	Happy			Sad			Normal			Average accuracy
	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1	
DLP [31]	62.25	66.76	64.42	53.04	82.96	64.90	79.70	56.20	65.92	65.02
Baseline [25]	65.13	80.53	72.01	82.24	88.86	85.42	91.21	72.18	80.59	79.57
SO-DL	52.75	92.34	67.14	93.91	85.52	89.52	97.03	73.13	83.40	81.31
SO-DLSE	59.62	91.73	72.27	93.54	89.22	91.33	96.04	73.98	83.58	83.13
VFESO-DLSE	78.5	93.18	85.21	96.27	85.07	90.33	88.86	86.71	87.78	87.90

Adaptive Moment Estimation (Adam) optimizer. It is worth pointing out that, to better measure the availability of the BabyExp dataset and the accuracy of the results, we report total accuracy, per class precision, per class recall, and per class F1-measure as the evaluation metrics here.

The last experiment is to verify the experimental results if the data are not equalized by CycleGAN. Table 5 shows the results of this experiment. The original dataset contains 7,600 pictures, including 2,502 happy images, 4,028 normal images, and 1,070 sad images. In order to objectively measure the performance, it is divided into training and test sets. The test set contains 1,522 images, and the remaining 6,078 images are used as the training set. We choose two methods with better experimental results in the second experiment: SO-DLSE and VFESO-DLSE. Experimental settings, parameter settings, and the number of iterations are the same as those in the second experiment above.

3. Results

Table 3 shows the experimental results of adult expression recognition models trained on the adult dataset and tested on the adult and BabyExp datasets. As we can see, the performance of these methods on the BabyExp is significantly lower than that on the adult dataset SFEW2.0, 54.45% on SFEW2.0 vs. 39.7% on BabyExp and 58.14% on SFEW2.0 vs. 40.78% on BabyExp, indicating that baby faces are greatly different from the adult faces, and it is important for developing facial expression recognition approaches for baby images.

The overall expression recognition performance of the proposed different experiments trained from scratch on the BabyExp dataset is shown in Table 4. From the results, we have the following observations: firstly, we can clearly see that the accuracy of DLP and baseline methods when trained and tested from scratch on the BabyExp dataset has greatly improved, 39.7% to 65.02% and 40.78% to 79.57%, compared with that trained on adult dataset SFEW2.0, once again indicating that baby faces are greatly different from the adult faces. Secondly, our proposed method VFESO-DLSE achieves the best result, 87.90%, which is about 4.8% greater than SO-DLSE showing that VGG16 is better than other CNN methods to extract features. From the results of baseline, SO-DL, and SO-DLSE, we can see distance loss and SE can achieve an improvement about 1.8%. The purpose of the distance loss is to learn lower changes between the same classes and higher distances between different classes, and the SE block can automatically obtain the importance of each feature channel through learning. Thirdly, from the results, it

Table 5: The expression recognition performance of original data which are not equalized by CycleGAN (trained from scratch).

Models	Happy			Sad			Normal			Average accuracy
	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1	
SO-DLSE	53.60	47.35	50.28	0.0	0.0	0.0	77.23	65.27	70.75	58.61
VFESO-DLSE	56.40	84.18	67.54	38.79	76.14	51.39	94.68	70.96	81.12	74.24

is obviously shown that the recall, precision, and F1-measure can further confirm the reliability of our results and the validity of our method.

The expression recognition performance of original data which are not equalized by CycleGAN can be seen in Table 5. We have two observations of the facial expression recognition on BabyExp. Firstly, we can easily see that two methods, SO-DLSE and VFESO-DLSE, have achieved 58.61% and 74.24% on the original data, in which both are still lower than 83.13% and 87.90% on BabyExp equalized by CycleGAN from Table 4. Secondly, even though these two methods have achieved higher accuracy, the recall rate and F1-measure are not very high, especially for the sad expression; this is because the distribution of expressions is unbalanced, and models trained using imbalance original facial expression samples have poor generalization ability and are prone to overfitting. Even in the SO-DLSE method, the recall, precision, and F1-score values of sad expressions are all 0, while the VFESO-DLSE method obtained 38.79%, 76.14%, and 51.39% in recall, precision, and F1-score, respectively, which also shows on the one hand that VGG16 is better than other CNN methods to extract features. On the other hand, it shows that we need to perform data augmentation to promote data balance and facilitate the use of deep learning methods for experiments, which validates the importance of CycleGAN for data equalization. This conclusion can also be drawn from the experimental results in Table 4.

4. Discussion

Facial expression recognition (FER) has always been a challenging topic in computer vision. Researchers usually aim to build a system that can identify different expressions in the images automatically [33]. Research on facial expression recognition relies heavily on an adequate dataset of facial expressions. However, due to the inherent nature of facial expressions and the difficulty of obtaining them, there are currently only a limited number of publicly available databases, which provide a sufficient number of facial images and are tagged with accurate facial expression information. Table 1 shows the summary of the existing image databases with the number of images, number of subjects, expression distribution, data size, and released years.

However, there are several limitations for these datasets. Most of the existing works and datasets [7, 8] focus on analyzing adult faces, which ignore how to analyze facial expressions from baby facial images. Recently, researchers have realized the importance of children facial expressions in order to study developmentally the interpretation of these expression datasets. For example, the new NIMH Children's Emotional Face Picture Collection (NIMH-ChEFS) contains photos of children aged 10–17 [12], the Radboud Faces Database includes photos of 8- to 12-year-olds [13], and the CAFE set features photographs of 2- to 8-year-old children [14]. Although these new datasets give researchers the option to use a sample of children aged 2–17 years, there have been no datasets that include younger children to date. On the contrary, all the datasets mentioned above for children facial expressions have only a small number of images, which are not suitable for training CNN models. In addition, these datasets contain posed expressions in the lab-controlled environment, not spontaneous or natural facial expressions.

5. Conclusions

In this paper, to address the aforementioned issues, we propose a new image dataset with expression labels of baby faces for automatic facial expression recognition. Our dataset, which we call the BabyExp dataset, contains more than 12,000 images from babies younger than two years old showing spontaneous expressions in an uncontrolled environment. Each face image is annotated with one of three facial expressions (i.e., happy, sad, and normal). It is complementary to the existing adult face dataset and can shed some light on exploring baby face analysis, and it will enable the academic research community to study baby faces in a manner comparable to the vast literature that relies heavily on adult faces.

As a result, our novel dataset will become an important milestone for human expression researchers. This dataset will be an important resource for the computer vision community to benchmark and compare results. We further evaluate state-of-the-art adult face analysis methods on BabyExp, which indicate that adult facial expression recognition methods are not suitable for baby facial expression recognition, and new methods are necessary to be developed to approach baby face recognition. Besides, we have also proposed a baseline for automatic expression recognition for babies based on deep learning. We conduct several experiments and report the baseline performances of the BabyExp dataset. The proposed baseline CNN architecture achieves an average classification accuracy of 87.90% on the BabyExp dataset. The performance of these methods on the BabyExp dataset is significantly lower than that on the other datasets, indicating that baby face facial images are greatly different from the adult faces, and it is important for the community to develop facial expression recognition approaches for babies.

We hope that the release of the BabyExp dataset will encourage more research works on the real-world children expression recognition, and it will be a useful benchmark

resource for researchers to validate their facial expression analysis algorithms in challenge conditions. We will collect more data and assign more specific facial expression labels (i.e., crying and laughing) to each image in order to extend the dataset. And we will continue to explore methods to achieve better performance for baby facial expression recognition in the future.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] M. Pantic and L. J. M. Rothkrantz, "Automatic analysis of facial expressions: the state of the art," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1424–1445, 2000.

[2] A. Di Domenico, R. Palumbo, N. Mammarella, and B. Fairfield, "Aging and emotional expressions: is there a positivity bias during dynamic emotion recognition?" *Frontiers in Psychology*, vol. 6, p. 1130, 2015.

[3] M. Altamura, F. A. Padalino, E. Stella et al., "Facial emotion recognition in bipolar disorder and healthy aging," *The Journal of Nervous and Mental Disease*, vol. 204, no. 3, pp. 188–193, 2016.

[4] R. Palumbo, R. B. Adams Jr, U. Hess, R. E. Kleck, and L. Zebrowitz, "Age and gender differences in facial attractiveness, but not emotion resemblance, contribute to age and gender stereotypes," *Frontiers in Psychology*, vol. 8, p. 1704, 2017.

[5] H. Ding, S. K. Zhou, and R. Chellappa, "Facenet2expnet: regularizing a deep face recognition net for expression recognition," in *Proceedings of the 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pp. 118–126, IEEE, Washington, DC, USA, June 2017.

[6] Y. Fan, J. C. Lam, and V. O. K. Li, "Multi-region ensemble convolutional neural network for facial expression recognition," in *Proceedings of the 27th International Conference on Artificial Neural Networks*, pp. 84–94, Springer, Rhodes, Greece, October 2018.

[7] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Static facial expression analysis in tough conditions: data, evaluation protocol and benchmark," in *Proceedings of the 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pp. 2106–2112, IEEE, Barcelona, Spain, November 2011.

[8] 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 *Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, pp. 94–101, IEEE, San Francisco, CA, USA, August 2010.

[9] H. Yang, Z. Zhang, and L. Yin, "Identity-adaptive facial expression recognition through expression regeneration using conditional generative adversarial networks," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pp. 294–301, IEEE, Xi'an, China, June 2018.

[10] C. M. Kuo, S. H. Lai, and M. Sarkis, "A compact deep learning model for robust facial expression recognition," in *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2121–2129, Salt Lake City, UT, USA, December 2018.

[11] P. D. M. Fernandez, F. A. G. Peña, T. I. Ren, and A. Cunha, "FERAtt: facial expression recognition with attention net," in *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Long Beach, CA, USA, June 2019.

[12] H. L. Egger, D. S. Pine, E. Nelson et al., "The NIMH child emotional faces picture set (NIMH-ChEFS): a new set of childrens̈ facial emotion stimuli," *International Journal of Methods in Psychiatric Research*, vol. 20, no. 3, pp. 145–156, 2011.

[13] O. Langner, R. Dotsch, G. Bijlstra, D. H. J. Wigboldus, S. T. Hawk, and A. Van Knippenberg, "Presentation and validation of the radboud faces database," *Cognition & Emotion*, vol. 24, no. 8, pp. 1377–1388, 2010.

[14] V. LoBue and C. Thrasher, "The child affective facial expression (CAFE) set: validity and reliability from untrained adults," *Frontiers in Psychology*, vol. 5, p. 1532, 2015.

[15] D. E. King, "A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, no. 60, pp. 1755–1758, 2009.

[16] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[17] I. Goodfellow, J. Pouget-Abadie, M. Mirza et al., "Generative adversarial nets," *Advances in Neural Information Processing Systems*, pp. 2672–2680, 2014.

[18] J. Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2223–2232, Venice, Italy, October 2017.

[19] M. Xia, X. Zhang, W. a. Liu, L. Weng, and Y. Xu, "Multi-stage feature constraints learning for age estimation," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 2417–2428, 2020.

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, no. 2, pp. 1097–1105, 2012.

[21] L. Weng, L. Wang, M. Xia, H. Shen, J. Liu, and Y. Xu, "Desert classification based on a multi-scale residual network with an attention mechanism," *Geosciences Journal*, pp. 1–13, 2020.

[22] M. Xia, W. Liu, Y. Xu, K. Wang, and X. Zhang, "Dilated residual attention network for load disaggregation," *Neural Computing and Applications*, vol. 31, no. 12, pp. 8931–8953, 2019.

[23] M. Xia, J. Qian, X. Zhang, J. Liu, and Y. Xu, "River segmentation based on separable attention residual network," *Journal of Applied Remote Sensing*, vol. 14, no. 3, Article ID 032602, 2019.

[24] K. Yu and M. Salzmann, "Second-order convolutional neural networks," 2017, http://arxiv.org/abs/1703.06817.

[25] D. Acharya, Z. Huang, D. Pani Paudel, and L. Van Gool, "Covariance pooling for facial expression recognition," in *Proceedings of the 2018 The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPRW)*, Salt Lake City, UT, USA, June 2018.

[26] J. Donahue, Y. Jia, O. Vinyals et al., "Decaf: a deep convolutional activation feature for generic visual recognition," *Proceedings of the 31st International Conference on Machine Learning*, vol. 32, no. 1, pp. 647–655, 2014.

[27] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: an astounding baseline for recognition," in *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 806–813, Columbus, OH, USA, June 2014.

[28] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, http://arxiv.org/abs/1409.1556v6.

[29] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141, Salt Lake City, UT, USA, June 2018.

[30] M. Abadi, A. Agarwal, P. Barham et al., "Large-scale machine learning on heterogeneous distributed systems," 2016, http://arxiv.org/abs/1603.04467.

[31] S. Li and W. Deng, "Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 356–370, 2018.

[32] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Proceedings of the European Conference on Computer Vision*, pp. 499–515, Springer, Amsterdam, The Netherlands, October 2016.

[33] S. Xie, H. Hu, and Y. Wu, "Deep multi-path convolutional neural network joint with salient region attention for facial expression recognition," *Pattern Recognition*, vol. 92, pp. 177–191, 2019.