Lightweight Facial Expression Recognition Based on Label Distribution Learning (Postprint)

Authors: Liu Jin, Luo Xiaoshu, Xu Zhaoxing

Date: 2022-04-07T15:01:57Z

Abstract

To address the issues of insufficient feature extraction for facial expressions and inadequate generalization capability of lightweight networks in complex environments, as well as ambiguous expressions caused by single-label datasets' inability to effectively describe complex emotional tendencies, a facial expression recognition method combining improved ShuffleNet with label distribution learning is proposed. Without substantially increasing computational complexity, a novel output module is designed to improve the ShuffleNet model to prevent overfitting; to enhance the model' s capacity for extracting important local detail features from facial expression images, a parallel deep convolutional residual module is designed, enabling the fusion of local and global features. To reduce the adverse impact of ambiguous expressions on recognition performance, the label distribution learning method is employed to fully utilize the dataset' s inherent information to generate label distributions without introducing additional information, and the improved ShuffleNet model is retrained. Experimental results show that the method achieves accuracies of 87.15%, 62.05%, and 58.49% on the RAF-DB, AffectNet-7, and AffectNet-8 datasets, respectively, while maintaining both the number of parameters and computational cost at low levels, which facilitates its application in practical production scenarios.

Full Text

Preamble

Research on Lightweight Facial Expression Recognition Based on Label Distribution Learning

Liu Jin[1], Luo Xiaoshu[1]†, Xu Zhaoxing[2]
(1. School of Electronic Engineering, Guangxi Normal University, Guilin, Guangxi 541000, China;

2. School of Big Data, Jiangxi Institute of Clothing Technology, Nanchang 330000, China)

Abstract: To address the problems of insufficient feature extraction for facial expressions in complex environments, limited generalization capability of lightweight networks, and the inability of single-label datasets to effectively describe ambiguous expressions caused by complex emotional tendencies, this paper proposes a facial expression recognition method that combines an improved ShuffleNet with label distribution learning. Without substantially increasing computational complexity, a new output module was designed to improve the ShuffleNet model to avoid overfitting. To enhance the model' s ability to extract important local detail features from facial expression images, a parallel depthwise convolution residual module was designed to achieve fusion of local and global features. To reduce the adverse effects of ambiguous expressions on recognition performance, the label distribution learning method was employed to fully utilize the original information of the dataset to generate label distributions without introducing additional information, and the improved ShuffleNet model was retrained. Experimental results demonstrate that the proposed method achieves accuracies of 87.15%, 62.05%, and 58.49% on the RAF-DB, AffectNet-7, and AffectNet-8 datasets, respectively, while maintaining low parameter count and computational cost, which facilitates its application in practical production environments.

Keywords: facial expression recognition; lightweight; label distribution learning; ambiguous expressions; depthwise separable convolution

0 Introduction

Since ancient times, "observing countenance" has been an important basis for comprehensively analyzing psychological activities. As stated in *The Analects of Confucius· Yan Yuan*: "The accomplished person is upright and loves righteousness, observes words and watches countenance, and considers how to humble themselves." Recognizing facial expressions to observe countenance can provide auxiliary structured information for individuals appearing in a scene. Therefore, facial expression recognition (FER) has extensive applications in affective computing, human-computer interaction, driver fatigue detection, teaching effectiveness evaluation, and many other fields [1,2]. In 1978, Ekman et al. [3] first defined six basic facial expressions in their cross-cultural research: happiness, sadness, anger, fear, disgust, and surprise. These basic emotions can be perceived, recognized, and shared by people from different cultural backgrounds.

In recent years, with the rapid development of deep learning in computer vision, it has also been successfully applied to facial expression recognition and has achieved significant progress. While deep learning technology improves expression recognition accuracy, it also leads to a sharp increase in the number of parameters and FLOPs. Although larger and deeper network models achieve

better performance, they also require higher hardware configurations during operation. However, in practical production and application environments, device configuration levels are often constrained by cost, and excessively high configuration requirements are not conducive to the practical deployment of models. Therefore, in the field of facial expression recognition, in addition to improving recognition accuracy, consideration should also be given to how to compress the computational overhead of models so that they can operate normally on small embedded devices with limited performance.

In 1980, psychologist Plutchik et al. [4] demonstrated that most human emotions are composed of basic facial expressions. In the real world, a static facial expression image is often composed of basic emotions of different intensities with complex emotional intentions, yet the expression image corresponds to only a single label. Due to the existence of such ambiguous expressions, the effectiveness of expression recognition is severely limited. Using label distribution learning (LDL) to address the problem that single labels cannot effectively describe complex emotional tendencies can further improve the recognition performance of FER models. Additionally, label distribution learning can alleviate noise problems caused by the subjectivity of dataset annotators and the ambiguity of expression images [5].

To address these two issues, this paper constructs a depthwise separable convolution residual module based on the lightweight network ShuffleNet, which can better extract features from key detail areas such as eyes and mouth in facial expression images without substantially increasing computational overhead. During training, the LDL method is used to generate label distributions, which helps improve the model' s discriminative ability for different expressions. This approach proposes a lightweight facial expression recognition method based on label distribution learning (LFER-LDL). The proposed method is experimentally validated on the RAF-DB [6] and AffectNet-7 [7] datasets, and the results demonstrate that the method achieves better recognition performance compared with some recently proposed expression recognition methods while maintaining low computational overhead.

1 Proposed Method

The facial expression recognition model proposed in this paper mainly consists of two parts: an improved ShuffleNet network and label distribution learning (LDL). The network structure is shown in Figure 1. The backbone network is ShuffleNet-V2 [8], which comprises Conv1, Stage2, Stage3, Stage4, and Conv5. To avoid overfitting and make the model more robust, this paper designs a new output module to replace the fully connected layer of the original network. To enhance the network' s ability to extract local detail features, a parallel depthwise convolution residual module (PDWRes) is designed without substantially increasing additional computational overhead. Based on Plutchik et al.' s research [4], to reduce the adverse effects of ambiguous expressions, the dataset itself is utilized to generate label distributions without using additional informa-

tion (right branch in Figure 1). The improved ShuffleNet network (left branch in Figure 1) has output channel numbers of 29, 116, 232, 464, and 1024 from Conv1 to Conv5 layers, respectively, and finally obtains seven-class facial expression recognition output through the Softmax layer.

1.1 Improved ShuffleNet Model

Deep convolutional neural networks (CNNs) such as ResNet and VGG can achieve high classification accuracy for expression images, but the computational complexity of these models also increases accordingly. Overly complex networks cannot meet the requirements of embedded device scenarios, and some mobile devices also require small models that are both fast and accurate. To satisfy these requirements, Ma N et al. [8] proposed the lightweight neural network ShuffleNet-V2, which can effectively balance the relationship between recognition accuracy and computational speed. ShuffleNet-V2 Unit primarily uses 1×1 pointwise convolution (PWConv) and depthwise convolution (DWConv), and performs Channel Shuffle operations on channel information within different feature groups to achieve information fusion between different groups.

Lin M et al. [9] demonstrated that the fully connected layer accounts for the largest proportion of parameters in CNN models. Although the fully connected layer can compress the dimension of feature maps and input them into the softmax layer to ultimately obtain seven-class facial expression images, this causes overfitting and is not conducive to enhancing the model' s generalization capability. Therefore, this paper designs an improved output module to replace the fully connected layer output module in the backbone network ShuffleNet-V2. The improved output module is shown in Figure 2.

The improved output module mainly consists of improved depthwise separable convolutions, similar to the pointwise and depthwise convolutions in the backbone network. The channel correlation and spatial correlation of depthwise separable convolution layers are decouplable [10]. Compared with standard convolution, depthwise separable convolution modules can further extract facial expression features without introducing a large number of parameters. When a convolution kernel of size acts on an input feature matrix of size , with input and output channel numbers of and respectively, the parameter count for standard convolution is while the parameter count for depthwise separable convolution is . The parameter count of depthwise separable convolution is only of that of standard convolution.

To prevent gradient vanishing, enhance the model' s non-linear capability, and reduce overfitting, ReLu activation functions are used after depthwise separable convolutions. Although ReLu enables faster backpropagation, neurons with input less than or equal to 0 will be suppressed, causing weights to be unable to update, which affects the final expression of the entire model. This paper improves the depthwise separable convolution module by replacing the ReLu activation function with the Mish activation function. The Mish activation

function formula is . The Mish activation function curve is shown in Figure 3. It retains a certain gradient flow for negative values, unlike the hard zero boundary in ReLu, which facilitates the flow of feature information. Additionally, every point on the Mish curve is smooth, which allows better information to penetrate deep into the neural network, thereby achieving better recognition accuracy and generalization.

1.2 Parallel Depthwise Convolution Residual Module Design

Facial expression recognition is often related to local detail features. For example, eyebrows, eyes, mouth, and other areas can more easily express different emotions, and the human eye also tends to focus on these regions when recognizing expressions. Therefore, to enable the network to effectively learn local detail features, this paper designs a parallel depthwise convolution residual module (PDWRes). By extracting features from local regions and supplementing them to the backbone network in the form of residual structures, the fusion of local and global features is achieved, making the network pay more attention to important features in facial expression images. The PDWRes module structure is shown in Figure 4.

For an input RGB facial expression image of size 224×224, after passing through the bottom Conv1, a feature map is obtained, where . Inspired by recent Transformer models [11], the feature map is divided into two equal parts horizontally and vertically to obtain four regional feature maps of facial expressions . Each small feature map sequentially undergoes two 3×3 DWConv operations to obtain detail feature maps of different facial regions. As described in Section 1.1, to avoid introducing a large number of computational parameters, only depthwise convolutions are used here for feature extraction. To accelerate model convergence, batch normalization (BN) is used after each depthwise convolution. To enhance model sparsity and reduce redundancy, ReLu6 activation function is used simultaneously after BN. ReLu6 is defined as follows:

The ReLu6 activation function sets the upper limit of the linear part of the ReLu function to 6, which is beneficial for achieving better numerical resolution on low-precision mobile devices and enhancing model stability.

Finally, the four regional feature maps are concatenated along the horizontal and vertical directions to obtain the complete local feature map . This is supplemented to the global feature after Stage2, and the global-local feature fusion expression is . As the network depth increases, the feature maps become smaller, which is not conducive to local feature extraction by the PDWRes module. Therefore, to minimize the introduction of additional computational overhead, the PDWRes module is only used in the Stage2 phase.

1.3 Label Distribution Learning

The annotation of facial expression images often requires significant human and material resources, and the emotional distribution is difficult to obtain, which

leads to ambiguous expressions and is not conducive to the classification of expression images. To compensate for the insufficient information of single labels during expression classification, this paper uses the label distribution learning method to generate expression distributions, as shown in the right branch of Figure 1. Its backbone network is ResNet-50. Different single-label facial expression datasets are pre-trained using this label distribution learning method to collect the overall distribution of the facial expression dataset. The generated data label distributions are then used to retrain the improved ShuffleNet network.

Given a facial expression image with label , where represents the number of expression image categories, label distribution learning will collect the distribution of expression images in the dataset . After passing through the fully connected layer (FC) of ResNet-50, is obtained. Label distribution learning finally uses a Softmax layer as output. The conditional probability that expression image belongs to category calculated by Softmax is .

To facilitate gradient backpropagation, this paper uses KL divergence to measure the difference between the prediction output of the improved ShuffleNet model and the label distribution obtained by LDL. KL divergence is non-negative, which satisfies the characteristics of deep learning gradient descent methods. However, due to its asymmetry, this paper uses the label distribution obtained by LDL as the true distribution , and the output of the improved ShuffleNet model as . Thus, the KL divergence for a sample size of can be written as .

Label distribution learning and KL divergence are only used during training to help the improved ShuffleNet network better learn the distribution and discrimination of facial expressions in the dataset. During testing, only the maximum value of the output probability from the improved ShuffleNet model' s softmax layer is used as the network output.

During the testing phase, combo loss is used as the loss function, which consists of a weighted sum of improved cross-entropy (CE loss) and dice loss. To control the regularization degree of false positives (FP) and false negatives (FN) for different datasets and correct network learning, binary cross-entropy is generalized to multi-class problems, and its output is the average of multiple binary cross-entropies. Dice Loss is mainly used to handle class imbalance problems in datasets and reduce model overfitting on easily classified expressions. Combo loss can be written as:

where and represent the true value and predicted value, respectively. Hyperparameter balances the weight of combo loss and improved cross-entropy. Hyperparameter controls the regularization degree of FP and FN, which is adjusted according to different datasets during experiments. To avoid division by zero, .

2 Experiments

2.1 Dataset Introduction

The experiments in this paper are conducted and evaluated on the large-scale facial expression datasets RAF-DB [6] and AffectNet [7]. Both RAF-DB and AffectNet-7 have seven categories of expression labels: sadness, surprise, disgust, fear, happiness, anger, and neutral. The AffectNet-8 dataset adds contempt expression on this basis, with eight categories of expression labels.

The RAF-DB dataset is the Real-world Affective Faces Database, containing 15,339 seven-class expression images. Each image is independently annotated by 40 people, divided into 12,271 training images and 3,068 test images. These expression images are affected by occlusion, pose, lighting conditions, and other factors, showing significant diversity and practical application value.

AffectNet is the largest facial expression dataset to date, containing over 1 million facial images from the Internet obtained by searching emotion labels through different search engines. Approximately half (440,000) of the images are annotated with 11 expression categories. This paper uses 290,000 manually labeled expression images from the AffectNet dataset as the training set, with 3,500 test images in AffectNet-7 and 4,000 test images in AffectNet-8. Figure 5 shows sample expression images from the RAF-DB and AffectNet-7 datasets.

2.2 Experimental Environment and Data Preprocessing

All experiments in this paper are completed under the Ubuntu 16.04 system, implemented based on the deep learning framework PyTorch 1.1 and Python 3.7 interpreter. The hardware environment includes an E5-2637 v4 CPU, NVIDIA GeForce GTX 1080Ti GPU with 11GB memory, and CUDA 10.2 acceleration library.

In the RAF-DB and AffectNet datasets collected in real-world scenarios, the size, angle, and pose of faces in expression images vary, which is not conducive to model learning. Therefore, RetinaFace [12] is used for face detection and alignment. To optimize model learning efficiency, the proposed method is pre-trained on the MS-Celeb-1M face dataset. To avoid overfitting, all expression images in the RAF-DB and AffectNet datasets are resized to 224×224 and randomly horizontally flipped with a probability of 0.5.

2.3 Experimental Settings

This paper uses Stochastic Gradient Descent (SGD) for training, with an initial learning rate of 0.01, momentum of 0.9, and weight decay . The model is trained for 120 epochs on both RAF-DB and AffectNet datasets. Due to differences in sample sizes across datasets, the batch size is 32 on the RAF-DB dataset, with the learning rate decaying by a factor of 0.1 every 30 epochs. On the AffectNet dataset, the batch size is 64, with the learning rate decaying by a factor of 0.1

every 10 epochs. Additionally, the AffectNet training set is imbalanced, but the test set is balanced, so a balanced sampling strategy is employed.

2.4 Experimental Results and Analysis

To verify the effectiveness of the proposed method and measure the model' s computational complexity, ShuffleNet-V2 is used as the backbone network, its output layer is improved, the PDWRes module is added, and label distribution learning is introduced. Experiments are conducted on the large-scale facial expression datasets RAF-DB and AffectNet, and the recognition accuracy and computational complexity are compared with other methods.

2.4.1 Influence of Balance Coefficient on Classification Performance This experiment investigates the influence of the balance coefficient in the Combo Loss function on recognition accuracy across different facial expression datasets. The balance coefficient controls the weight of Dice Loss on . During experiments, Combo Loss and improved cross-entropy are assigned equal weights, i.e., . The balance coefficient controls the penalty degree of improved cross-entropy on FP and FN. When is less than 0.5, has greater weight, and FP receives more penalty than FN, and vice versa. During experiments, is 取值 from 0 to 1 in steps of 0.1.

The experimental results on the RAF-DB dataset are shown in Figure 6. The expression recognition accuracy first increases and then decreases as the balance coefficient increases, reaching the highest recognition accuracy of 87.15% when is 0.2. When is less than 0.2, the model' s recognition accuracy is insufficient, and when is greater than 0.2, the model's recognition accuracy begins to decline. This indicates that for the RAF-DB dataset, a larger penalty for false positive sample images is needed to assist model learning in achieving better recognition accuracy.

2.4.2 RAF-DB Experimental Results Figure 8 shows the training and testing accuracy curves and loss function curves on the RAF-DB expression dataset. For clear display in the same coordinate system, the loss function curve is magnified by 30 times. As can be seen from the figure, the model basically converges at the 35th epoch, and the final recognition accuracy difference between the training set and test set is small, thanks to the improved output module of the ShuffleNet model, which avoids model overfitting. The model ultimately achieves a recognition accuracy of 87.15% on the RAF-DB dataset.

To further verify the effectiveness of the proposed model and measure its computational complexity, a comparison with other methods in recent literature is conducted on the RAF-DB expression dataset, as shown in Table 1. In terms of parameter count, the proposed method has only 1.26M parameters, which is far lower than the 134.29M of gACNN [14], and compared with methods with smaller parameter counts such as Separate Loss [15], RAN [16], and DDA Loss [18], the parameter count of the proposed method is only one-tenth of theirs,

significantly compressing the model' s parameter count. In terms of FLOPs, the proposed method' s 294.60M FLOPs represents a 98.09% reduction in computational cost compared with gACNN, and an 83.81% reduction compared with Separate Loss and DDA Loss methods, giving the proposed model lower complexity. In terms of accuracy, compared with recently proposed RAN and DDA Loss methods, the proposed method's accuracy is improved by 0.25%. Compared with IPA2LT [13], gACNN, Separate Loss, and LDL-ALSG [17], the proposed model' s recognition accuracy is improved by 0.38%, 2.08%, 0.77%, and 1.62%, respectively. Since labels in datasets may contain annotation errors, Wang et al. [19] proposed the self-curing network SCN, which corrects network learning through regularization ranking and relabeling operations, achieving 87.03% accuracy on the RAF-DB dataset. Compared with SCN, the proposed method' s accuracy is improved by 0.12%, while the parameter count and FLOPs are compressed by 10 times and 6 times, respectively, verifying the effectiveness of the proposed method. It can be seen that the proposed method maintains good recognition accuracy while keeping parameter count and computational cost low, which is beneficial for practical application of the model in production environments.

2.4.3 AffectNet Experimental Results The experimental results on the AffectNet-7 dataset are shown in Figure 7. The recognition accuracy first decreases, then increases, and then decreases again as the balance coefficient increases, reaching the highest recognition accuracy of 62.05% when is 0.6. When is less than 0.6, the model accuracy first decreases and then increases, and when is greater than 0.6, the model' s recognition accuracy begins to decline significantly. For the AffectNet dataset, penalty for false negative sample images is required. Experimental results show that the balance coefficient has a significant impact on the network' s recognition performance, and the selection of balance coefficient is crucial under different datasets.

Figure 9 shows the training and testing accuracy curves and loss function curves on the AffectNet-7 expression dataset. Like the RAF-DB experiment, the loss function curve is magnified by the same factor. As can be seen from the figure, the model basically converges at the 15th epoch, demonstrating fast fitting speed, which benefits from the LDL module assisting model learning to achieve rapid and stable convergence, also facilitating the model' s operation on actual embedded devices and avoiding overfitting problems. The model ultimately achieves a recognition accuracy of 62.05% on the large-scale expression dataset AffectNet-7.

To verify the effectiveness of the proposed method on the AffectNet-7 dataset and its computational complexity, a comparison with other recent methods is conducted, as shown in Table 2. In terms of parameter count, the proposed method is only 0.8% of the 145M parameters of VGG Face [21], and compared with other methods, the parameter count of the proposed method is only 0.93%~51.01% of theirs, maintaining a low parameter count. In terms of

FLOPs, compared with VGG Face' s 15490.46M, the proposed method compresses computational cost by 98.1%, similar to the compression achieved with gACNN. Compared with seven other methods, the proposed method also compresses computational cost by 61.40%~94.8%. In terms of accuracy, compared with recently proposed VGG Face and LDL-ALSG, the proposed method' s recognition accuracy is improved by 2.05% and 2.70%, respectively. Compared with IPA2LT, gACNN, Separate Loss, IPFR, and FMPN methods, the proposed accuracy is improved by 4.74%, 3.27%, 3.16%, 4.65%, and 0.53%, respectively. Although the proposed method' s recognition accuracy is not as high as SNA-DFER and DDA Loss, the parameter counts of SNA-DFER and DDA Loss are 1.9 times and 8.8 times that of the proposed method, respectively, and their FLOPs are 2.5 times and 6.1 times that of the proposed method, respectively, which is not conducive to running the model on embedded devices with lower performance. Overall, the proposed method effectively reduces model complexity while maintaining high-level expression recognition performance, verifying its effectiveness and practicality.

To further verify the effectiveness of the proposed method on the AffectNet-8 dataset containing 8 emotion categories and evaluate its parameter count and FLOPs, a comparative analysis with other methods is conducted, as shown in Table 3. The proposed method achieves 58.49% accuracy on the AffectNet-8 dataset. In terms of parameter count, compared with Weighted-loss, VGGNet-Variant, and RAN methods, the parameter count of the proposed method is only 2.21%, 19.27%, and 11.26% of theirs, respectively. In terms of FLOPs, Weighted-loss' s FLOPs are approximately 2400 times that of the proposed method, while RAN and ESR-9 methods are 49 times and 3 times that of the proposed method, respectively. In terms of accuracy, compared with Weighted-loss, MobileNet-Variant, and VGGNet-Variant methods, the proposed method improves accuracy by 0.49%, 2.49%, and 0.49%, respectively. Although the proposed method' s accuracy is not as high as RAN and ESR-9, its FLOPs are far lower than both methods. The MobileNet-Variant method achieves good results in both parameter count and FLOPs, but its accuracy is about 2.5% lower than the proposed method. The VGGNet-Variant method also achieves lower FLOPs, but its performance in parameter count and accuracy is not as good as the proposed method. It can be seen that model computational complexity and model performance cannot be obtained simultaneously, but the proposed method still achieves good performance on the AffectNet-8 dataset while maintaining low parameter count and FLOPs.

As can be seen from Table 3, AffectNet-8 is a challenging facial expression dataset. There is also a certain difference in accuracy between the proposed method on AffectNet-7 and AffectNet-8 datasets. AffectNet-8 adds contempt expression on the basis of AffectNet-7. By observing the dataset, it is found that the contempt expression in the AffectNet-8 dataset contains a large number of images that are not of this expression, such as happiness, etc. The label noise caused by the subjectivity of annotators is not conducive to network learning. Figure 10 shows some images in the contempt expression that do not belong to

contempt.

2.4.4 Ablation Study The proposed method includes improvements to the output module, design of the parallel depthwise convolution residual module, and label distribution learning. To analyze the impact of different components on facial expression recognition performance, an ablation study is conducted on the RAF-DB dataset.

This section uses ShuffleNet as the baseline and sequentially adds the improved output module, parallel depthwise convolution residual module, and label distribution learning to analyze the impact of the three modules on recognition performance. The experimental results are shown in Table 4. By improving the output module to extract high-dimensional facial expression features, the recognition accuracy is improved by 0.47% compared with the ShuffleNet baseline network, with parameter count increased by 0.02M and FLOPs increased by 0.03M. This benefits from the depthwise separable convolution in the improved output module for further extraction of facial expression features, while the Mish activation function used also ensures the flow of feature information. The parallel depthwise convolution residual module obtains local region features through integrated depthwise convolutions, making the network pay more attention to subtle differences among different expressions. With parameter count increased by 0.01M and FLOPs increased by 3.09M, it achieves a 1.36% improvement compared with the baseline network. By fusing local features into global features, the model focuses more on discriminative features in facial expression images, a characteristic similar to the working principle of the human eye. Label distribution learning helps reduce the impact of ambiguous expressions without introducing additional parameters and FLOPs, ultimately achieving 87.15% accuracy, which is 5.13% higher than the original network. Facial expression images in the real world often have complex emotional intentions. Label distribution learning collects the distribution of expression images in the dataset to reduce the uncertainty of ambiguous expressions, which helps alleviate the problem of insufficient information from single labels, demonstrating the effectiveness of the proposed method.

Finally, to visualize the results of the proposed method, the trained network model is saved and used for facial expression recognition. Images are randomly selected from the Internet and partially from datasets for instance testing, with test results shown in Figure 11.

3 Conclusion

Facial expression recognition has extensive applications in many fields. However, in practical production environments, overly complex network models are not conducive to operation on devices with limited configurations. Therefore, this paper proposes a lightweight facial expression recognition method based on label distribution learning. From the perspective of feature extraction, this method

improves the traditional ShuffleNet network model and designs a parallel depthwise convolution residual module, which enhances the model' s ability to extract local detail features from facial expression images. In terms of training strategy, label distribution learning is used to solve the problem of ambiguous expressions caused by insufficient information from single labels. Finally, the influence of the balance coefficient in the Combo Loss function on different facial expression datasets is analyzed. Comparative experiments are conducted on the RAF-DB, AffectNet-7, and AffectNet-8 datasets. Experimental results demonstrate that the proposed method maintains high recognition accuracy while keeping parameter count and FLOPs low, showing strong practicality.

Deep learning models in facial expression recognition research often require large amounts of annotated data, which not only incurs expensive annotation costs but may also introduce label noise from subjective factors. Therefore, future work will investigate semi-supervised or unsupervised learning for facial expression recognition.

References

[1] Yu M, Guo Z, Yu Y, et al. Spatiotemporal featuredescriptor for microexpression recognition using local cube binarypattern [J]. IEEE Access, 2019, 7: 214-225.

[2] Zheng Jian, Zheng Chi, Liu Hao, et al. Deep convolutional neural network fusing local feature and two-stage attention weight learning for facial expression recognition [J]. Application Research of Computers, 2021.

[3] Ekman P, Friesen W V. Facial Action Coding System (FACS): A Technique for the Measurement of Facial Actions [J]. Rivista Di Psichiatria, 1978, 47 (2): 126-38.

[4] Plutchik R. A general psychoevolutionary theory of emotion [M]. Theories of emotion. Academic press, 1980: 3-33.

[5] Chen S, Wang J, Chen Y, et al. Label distribution learning on auxiliary label space graphs for facial expression recognition [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 13984-13993.

[6] Li S, Deng W. Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition [J]. IEEE Trans on Image Processing, 2019, 28 (1): 356-370.

[7] Ali M, Behzad H, Mohammad H. Mahoor. AffectNet: A New Database for Facial Expression, Valence, and Arousal Computation in the Wild [J]. IEEE Trans on Affective Computing, 2017.

[8] Ma N, Zhang X, Zheng H T, et al. Shufflenet v2: Practical guidelines for efficient cnn architecture design [C]// Proceedings of the European conference on computer vision (ECCV). 2018: 116-131.

[9] Lin M, Chen Q, Yan S. Network in network [J]. arXiv preprint arXiv: 1312. 4400, 2013.

[10] Sifre L, Mallat S. Rotation, scaling and deformation invariant scattering for texture discrimination [C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2013: 1233-1240.

[11] Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows [J]. arXiv preprint arXiv: 2103. 14030, 2021.

[12] Deng J, Guo J, Ververas E, et al. Retinaface: Single-shot multi-level face localisation in the wild [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 5203-5212.

[13] Zeng J, Shan S, Chen X. Facial expression recognition with inconsistently annotated datasets [C]// Proceedings of the European conference on computer vision (ECCV). 2018: 222-237.

[14] Li Y, Zeng J, Shan S, et al. Occlusion aware facial expression recognition using CNN with attention mechanism [J]. IEEE Trans on Image Processing, 2018: 1-1.

[15] Li Y, Lu Y, Li J, et al. Separate loss for basic and compound facial expression recognition in the wild [C]// Asian Conference on Machine Learning. PMLR, 2019: 897-911.

[16] Wang K, Peng X, Yang J, et al. Region Attention Networks for Pose and Occlusion Robust Facial Expression Recognition [J]. IEEE Trans on Image Processing, 2020, PP (99): 1-1.

[17] Chen S, Wang J, Chen Y, et al. Label distribution learning on auxiliary label space graphs for facial expression recognition [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 13984-13993.

[18] Farzaneh A H, Qi X. Discriminant distribution-agnostic loss for facial expression recognition in the wild [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2020: 406-407.

[19] Wang K, Peng X, Yang J, et al. Suppressing uncertainties for large-scale facial expression recognition [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 6897-6906.

[20] Wang C, Wang S, Liang G. Identity-and pose-robust facial expression recognition through adversarial feature learning [C]// Proceedings of the 27th ACM International Conference on Multimedia. 2019: 238-246.

[21] Kollias D, Cheng S, Ververas E, et al. Deep neural network augmentation: Generating faces for affect analysis [J]. arXiv preprint arXiv: 1811. 05027, 2018.

[22] Chen Y, Wang J, Chen S, et al. Facial motion prior networks for facial

expression recognition [C]// 2019 IEEE Visual Communications and Image Processing (VCIP). IEEE, 2019: 1-4.

[23] Fu Y, Wu X, Li X, et al. Semantic neighborhood-aware deep facial expression recognition [J]. IEEE Transactions on Image Processing, 2020, 29: 6535-6548.

[24] Hewitt C, Gunes H. Cnn-based facial affect analysis on mobile devices [J]. arXiv preprint arXiv: 1807. 08775, 2018.

[25] Siqueira H, Magg S, Wermter S. Efficient facial feature learning with wide ensemble-based convolutional neural networks [C]// Proceedings of the AAAI conference on artificial intelligence. 2020, 34 (04): 5800-5807.

Note: Figure translations are in progress. See original paper for figures.