Combating Label Ambiguity with Smooth Learning for Facial Expression Recognition

Yifan Chen[1], Zide Liu[2], Xuna Wang[2], Shengnan Xue[3], Jiahui Yu[3], and Zhaojie Ju[1(✉)]

[1] School of Computing, University of Portsmouth, Portsmouth PO13HE, UK
zhaojie.ju@port.ac.uk
[2] School of Automation and Electrical Engineering, Shenyang Ligong University, Shenyang 110000, China
[3] Zhejiang Univeristy, Hangzhou 310027, China
{3210105824,jiahui.yu}@zju.edu.cn

Abstract. Accurately learning facial expression recognition (FER) features using convolutional neural networks (CNNs) is a non-trivial task because of the presence of significant intra-class variability and inter-class similarity as well as the ambiguity of the expressions themselves. Deep metric learning (DML) methods, such as joint central loss and softmax loss optimization, have been adopted by many FER methods to improve the discriminative power of expression recognition models. However, equal supervision of all features with DML methods may include irrelevant features, which ultimately reduces the generalization ability of the learning algorithm. We propose the Attentive Cascaded Network (ACD) method to enhance the discriminative power by adaptively selecting a subset of important feature elements. The proposed ACD integrates multiple feature extractors with smooth center loss to extract to discriminative features. The estimated weights adapt to the sparse representation of central loss to selectively achieve intra-class compactness and inter-class separation of relevant information in the embedding space. The proposed ACD approach is superior compared to state-of-the-art methods.

Keywords: Deep Metric Learning · Ambiguous Expressions · Facial Expression Recognition

1 Introduction

In the past few years, facial expression recognition has attracted increasing attention in the field of human-computer interaction [5,13,21,25]. Facial expressions

The authors would like to acknowledge the support from the National Natural Science Foundation of China (52075530), the AiBle project co-financed by the European Regional Development Fund, and the Zhejiang Provincial Natural Science Foundation of China (LQ23F030001).

H. Yang et al. (Eds.): ICIRA 2023, LNAI 14268, pp. 127–136, 2023.
https://doi.org/10.1007/978-981-99-6486-4_11

can be seen as reflecting a person's mental activity and mental state. With the rapid growth in the field of human-computer interaction, scientists have conducted a great deal of research to develop systems and robots that can automatically sense human feelings and states [24]. The ultimate goal is to sense human emotional states and interact with the user in the most natural way possible. This is a very complex and demanding task, as performing expression recognition in real-world conditions is not easy and straightforward. Facial expression recognition is significant in human-computer interaction. Although facial expression recognition has been studied and developed for many years, achieving accurate facial expression recognition is still challenging.

One of the main challenges of facial expression recognition is the labeling ambiguity problem. There are two reasons: one is the ambiguity of the expression itself, where some expressions are similar and difficult to distinguish. The other is the labeling ambiguity caused by different people, resulting in inconsistent labeling. For example, "happy" and "surprise" are similar and hard to distinguish. Moreover, the model may learn unuseful facial expression features instead of helpful information resulting in insufficient accuracy. Developing robust facial recognition systems is still a challenging task. Three elements primarily affect FER tasks based on deep learning techniques: data, models, and labels [2]. Researchers have made significant advances in models and data, but they need to pay more attention to labels. Xu et al. [22] suggested a Graph Laplacian Label Enhancement (GLLE) recover distribution from logical labels. However, the algorithm's rigid feature space topology assumptions make it unsuitable for big field datasets.

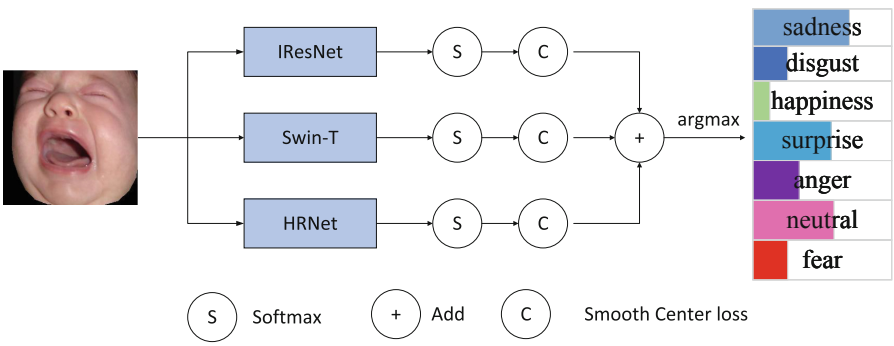

Fig. 1. They have different lateral connections (with or without skip connections), feature extract methods, and resolution streams (gradually decreasing or staying the same). Alternative architectures could be able to pick up different functionalities. We are ensembling these three backbones in the coarse net to get reliable prediction results and avoid over-fitting.

To solve this problem, we propose a cascade network to obtain more reliable features in different ways as shown in Fig. 1. Specifically, we train multiple models

based on various architectures and improve the whole performance using an ensemble. Finally, Joint optimization using softmax and smooth center loss.

The main contributions of our work can be summarized as follows:

- We propose the cascaded networks to address the label ambiguity problem in facial expression recognition.
- We propose the smooth center loss selectively achieves intra-class compactness and inter-class separation for the relevant information in the embedding space. Smooth center loss is jointly optimized with softmax loss and can be trained.

2 Related Work

Facial expression recognition is an important research topic in the field of computer vision and human-computer interaction. The earliest expression recognition methods were based on hand-crafted features [1,27]. Recently, deep learning methods have significantly advanced the development of facial expression recognition [26]. Some works [6,23] regard multi-branch networks to capture global and local features. A hybrid architecture combining CNN and Transformer has achieved state-of-the-art performance in several benchmarks to improve recognition generalization. Recently, several researchers [5,21] proposed extracting discriminative features through an attention mechanism, which was robust to occlusions.

Deep Metric Learning (DML) approaches constrain the embedding space to obtain well-discriminated deep features. Identity-aware convolutional neural network can simultaneously distinguish expression-related and identity-related features [18]. They employed contrast loss on depth features to combine features with similar labels and separate features with different labels. Similarly, Liu et al. [15] proposed the (N+M)-tuplet clusters loss function. By constructing a set of N-positive samples and a set of M-negative samples, the negative samples are encouraged to move away from the center of positive samples while the positive samples cluster around their respective centers. This integration improves intra-class compactness by leveraging the k-nearest neighbor algorithm for the local clustering of deep features. Furthermore, Farzaneh and Qi [8] proposed the discriminative distribution uncertainty loss, which in the case of class imbalance of the forward propagation process, regulates the Euclidean distance between the classes in the embedding space of the samples.

3 Methodology

This section briefly reviews the necessary preliminaries related to our work. We then introduce the two building blocks of our proposed Attentive Cascaded Network (ACN): the smooth center loss and the cascaded network. Finally, we discuss how ACN is trained and optimized.

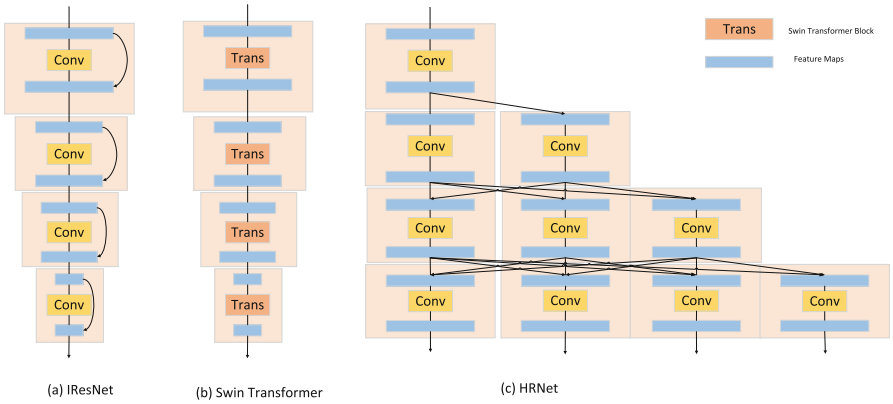

Fig. 2. An illustration of our collaborative methodology. Each backbone network's feature logits are extracted independently, and the softmax function generates the per-class confidential scores. The argmax function is used to generate the ultimate result after adding the scores.

3.1 Cascaded Network

Prior studies split expressions into seven basic categories to cover emotions common to humans. However, these seven categories of expressions are very similar, especially negative ones. Particularly, four negative expressions-anger, contempt, fear, and sadness-have comparable facial muscle motions. In contrast to positive sentiments, negative ones are more challenging to accurately forecast. The majority of in-the-wild facial expression datasets, on the other hand, are gathered from the Internet, where people typically share pleasant life experiences. Negative emotions are difficult to obtain in real scenarios, making the existing FER training dataset unbalanced regarding category distribution.

Specifically, we use two branches to predict positive expressions (happy, surprised, normal) and negative labels (anger, disgust, fear, and sadness). In this way, the low-frequency negative samples are combined in the negative samples, making the training dataset more balanced.

As shown in Fig. 2. We use the model ensemble strategy to our coarse net in order to further increase the robustness of our framework. The backbone networks are specifically HRNet, Swin-S, and IResNet-152. These architectures differ significantly from one another, as seen in Fig. 3, and each one extracts distinctive features. Pooling or path merging layers are used by IResNet (a) and Swin Transformer (b) to reduce the spatial resolution of feature maps, which lowers processing costs and broadens the perceptual field. To gain rich semantic features, HRNet (c) continues to be the high-resolution representation and exchanges features across resolutions. Unlike the other models, which employ the conventional convolution method, model (b) uses the attention mechanism with shifted windows to study relationships with other locations. (c) uses more connections between different resolutions to extract rich semantic features. These

different architectural designs can help different models learn different features and prevent the whole framework from overfitting to some noisy features.

3.2 Smooth Central Loss

Center loss, a widely used Deep Metric Learning technique, assesses how similar the deep features are to the class centers that correspond to them. The goal of center loss is to reduce the sum of squares between deep features and their corresponding class centers within each cluster, mathematically represented as shown below. Specifically, given a training minibatch of m samples,

$$\mathcal{L}_C = \frac{1}{2n} \sum_{i=1}^{n} \sum_{j=1}^{m} \|x_{ij} - c_{y_i j}\|_2^2 \tag{1}$$

where the center loss penalizes the Euclidean distance between a depth feature and its corresponding class center in the embedding space. The depth features are made to cluster at the class centers.

Not all elements in a feature vector are useful for classification. Therefore, we select only a subset of elements in the deep feature vector to help discriminate. Our goal is to filter out irrelevant features during the classification process, and we assign weights to the Euclidean distance in each dimension in Eq. 3 and develop a smooth central loss method as follows:

$$\mathcal{L}_{SC} = \frac{1}{2n} \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_{ij} \otimes \|x_{ij} - c_{y_i j}\|_2^2 \tag{2}$$

where $\otimes$ indicates element-wise multiplication and denotes the weight of the deep feature along the dimension in the embedding space. It should be noted that $\mathcal{L}_{SC}$ and $\mathcal{L}_C$ are the same if $\alpha_{ij} = 1$.

4 Experimental Settings and Results

In this section, we first present two publicly available FER datasets, the in-the-lab dataset ck+ [16] and the Real World Affective Facial Database (RAF-DB) [10]. Then, we conducted validation experiments on these two widely used facial expression recognition (FER) datasets to demonstrate the superior performance of our proposed Attentive Cascaded Network (ACN). Finally, we evaluated our method on the publicly available FER dataset compared to two baselines and various state-of-the-art methods.

4.1 Datasets

RAF-DB: The RAF-DB contains 12,271 training images and 3,068 images. It is a facial image obtained by crowdsourcing techniques and contains happy, sad, surprised, angry, fearful, disgusted, and neutral expressions. The dataset are acquired in an unconstrained setting offering a broad diversity across pose, gender, age, demography, and image quality.

CK+: A total of 123 individual subjects are represented by 593 video sequences in the Extended Cohn-Kanade (CK+) dataset. One of the seven expression classes-anger, contempt, disgust, fear, pleasure, sorrow, and surprise are assigned to 327 of these movies. Most facial expression classification methods employ the CK+ database, which is largely recognized as the most frequently used laboratory-controlled facial expression classification database available.

4.2 Implementation Details

Our experiments use the standard convolutional neural network (CNN) ResNet-18 as the backbone architecture. Before performing the expression recognition task, we pre-trained ResNet-18 on Imagenet, a face dataset containing 12 subtrees with 5247 synsets and 3.2 million images. We employ a typical Stochastic Gradient Descent (SGD) optimizer with weight decay of 5×10^{-4} and momentum of 0.9. We add new elements to the supplied photographs instantly by removing arbitrary crops. We utilize the supplied image's middle crop for testing. Crops with dimensions of 224× 224 are taken from input photos with dimensions of 256× 256.

We train ResNet-18 on the public dataset for 80 epochs with an initial learning rate of 0.01, decaying by a factor of 10 every 20 periods. The batch size is set to 128 for both datasets. The hyper-parameters α and λ are empirically set to 0.5 and 0.01, respectively. Our experiments use the PyTorch deep learning framework on an NVIDIA 1080Ti GPU with 8GB of V-RAM.

4.3 Recognition Results

Table 1 displays the results for RAF-DB, while Table 2 presents the results for CK+. Notably, the test set of RAF-DB is characterized by an unbalanced distribution. As a result, we provide average accuracy, computed as the mean of the diagonal values in the confusion matrix, and the standard accuracy, which encompasses all classes in RAF-DB.

Table 1. Performance of different methods on RAF-DB

Method	Accurancy
Gate-OSA [14]	86.32
gaCNN [12]	85.07
LDL-ALSG [6]	85.53
PAT-ResNet [4]	84.19
NAL [9]	84.22
Center Loss [11]	82.86
PAT-VGG [4]	83.83
ACD	**86.42**

Table 2. Performance of different methods on CK+

Method	Accurancy
FN2EN [7]	96.80
Center Loss [3]	92.26
DRADAP [17]	90.63
IL-CNN [3]	94.35
IDFERM [15]	98.35
Block-FerNet [20]	98.41
DeepEmotion [19]	90.63
ACD	**99.12**

As can be seen from Table 1, our ACN method outperforms the baseline method and other state-of-the-art methods, achieving 86.42% recognition accuracy on RAF-DB. In addition, the improvement of ACN over the two baseline methods is greater than the improvement of center loss over softmax loss. In other words, ACN significantly enhances the generalization ability of the model.

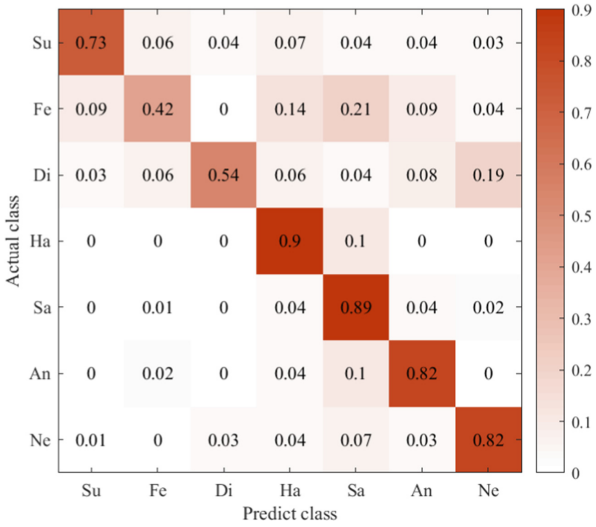

Fig. 3. The ACD framework in this paper: the diagonal line is the proportion of correctly identified and the non-diagonal line is the proportion of confused.

As shown in the Fig. 3 and Fig. 4, the confusion matrices obtained by the baseline approach and our proposed ACN framework are shown on the two FER datasets to analyze each category's recognition accuracy visually. Compared with softmax loss, ACN improves the recognition accuracy of all categories except surprise in the RAF-DB test set. The overall performance of ACN on RAF-DB is better because the recognition accuracy for surprise, fear, and disgust is significantly higher than that of central loss. We note that ACN outperforms the baseline approach on CK+ except for the anger category, while the recognition accuracy for the sadness and disgust categories is significantly higher than both baselines. Overall, ACN outperformed the baseline method for all classes in both RAF-DB and CK+.

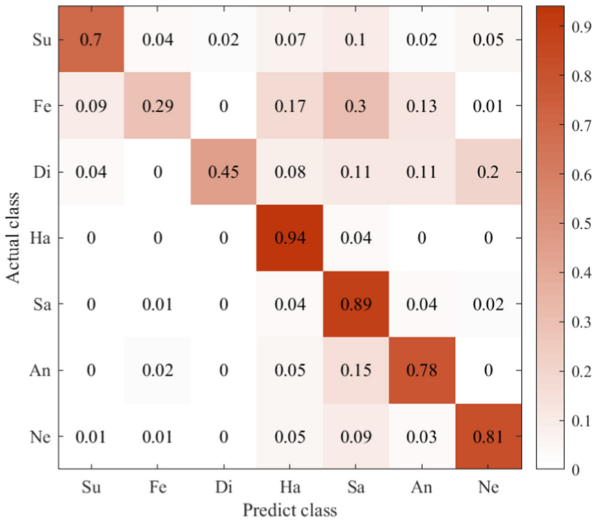

Fig. 4. The confusion matrix obtained from the baseline method (softmax loss)

5 Conclusions

This paper proposes an enhanced robustness approach called Attentive Cascaded Network (ACN). Our hybrid system uses smoothed central loss to enable the model to learn discriminative features that can distinguish between similar expressions. In addition, a cascaded network is proposed to address the label ambiguity problem. Our experimental results show that ACD outperforms other state-of-the-art methods on two publicly available FER datasets, namely RAF-DB and CK+.

ACD can easily be applied to other network models to solve other classification tasks and increase feature discrimination. In the future, we can extend the model to gesture and hand gesture recognition.

References

1. Amos, B., et al.: OpenFace: a general-purpose face recognition library with mobile applications. CMU School Comput. Sci. **6**(2) (2016)
2. Bagherinezhad, H., Horton, M., Rastegari, M., Farhadi, A.: Label refinery: improving ImageNet classification through label progression. arXiv preprint arXiv:1805.02641 (2018)
3. Cai, J., et al.: Island loss for learning discriminative features in facial expression recognition. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), pp. 302–309. IEEE (2018)
4. Cai, J., Meng, Z., Khan, A.S., Li, Z., O'Reilly, J., Tong, Y.: Probabilistic attribute tree in convolutional neural networks for facial expression recognition. arXiv preprint arXiv:1812.07067 (2018)

5. Chen, C., Crivelli, C., Garrod, O.G., Schyns, P.G., Fernández-Dols, J.M., Jack, R.E.: Distinct facial expressions represent pain and pleasure across cultures. Proc. Natl. Acad. Sci. **115**(43), E10013–E10021 (2018)
6. Chen, S., Wang, J., Chen, Y., Shi, Z., Geng, X., Rui, Y.: Label distribution learning on auxiliary label space graphs for facial expression recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13984–13993 (2020)
7. Ding, H., Zhou, S.K., Chellappa, R.: FaceNet2ExpNet: regularizing a deep face recognition net for expression recognition. In: 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), pp. 118–126. IEEE (2017)
8. Florea, C., Florea, L., Badea, M.S., Vertan, C., Racoviteanu, A.: Annealed label transfer for face expression recognition. In: BMVC, p. 104 (2019)
9. Goldberger, J., Ben-Reuven, E.: Training deep neural-networks using a noise adaptation layer (2016)
10. Li, S., Deng, W.: Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. IEEE Trans. Image Process. **28**(1), 356–370 (2018)
11. Li, S., Deng, W., Du, J.: Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2852–2861 (2017)
12. Li, Y., Zeng, J., Shan, S., Chen, X.: Occlusion aware facial expression recognition using CNN with attention mechanism. IEEE Trans. Image Process. **28**(5), 2439–2450 (2018)
13. Lin, Z., et al.: CAiRE: an empathetic neural chatbot. arXiv preprint arXiv:1907.12108 (2019)
14. Liu, H., Cai, H., Lin, Q., Li, X., Xiao, H.: Adaptive multilayer perceptual attention network for facial expression recognition. IEEE Trans. Circuits Syst. Video Technol. **32**(9), 6253–6266 (2022). https://doi.org/10.1109/TCSVT.2022.3165321
15. Liu, X., Kumar, B.V., Jia, P., You, J.: Hard negative generation for identity-disentangled facial expression recognition. Pattern Recogn. **88**, 1–12 (2019)
16. Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The extended Cohn-Kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, pp. 94–101. IEEE (2010)
17. Mandal, M., Verma, M., Mathur, S., Vipparthi, S.K., Murala, S., Kumar, D.K.: Regional adaptive affinitive patterns (RADAP) with logical operators for facial expression recognition. IET Image Proc. **13**(5), 850–861 (2019)
18. Meng, Z., Liu, P., Cai, J., Han, S., Tong, Y.: Identity-aware convolutional neural network for facial expression recognition. In: 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), pp. 558–565. IEEE (2017)
19. Minaee, S., Minaei, M., Abdolrashidi, A.: Deep-Emotion: facial expression recognition using attentional convolutional network. Sensors **21**(9), 3046 (2021)
20. Tang, Y., Zhang, X., Hu, X., Wang, S., Wang, H.: Facial expression recognition using frequency neural network. IEEE Trans. Image Process. **30**, 444–457 (2020)
21. Wells, L.J., Gillespie, S.M., Rotshtein, P.: Identification of emotional facial expressions: effects of expression, intensity, and sex on eye gaze. PLoS ONE **11**(12), e0168307 (2016)
22. Xu, N., Liu, Y.P., Geng, X.: Label enhancement for label distribution learning. IEEE Trans. Knowl. Data Eng. (2019)

23. Xu, N., Shu, J., Liu, Y.P., Geng, X.: Variational label enhancement. In: International Conference on Machine Learning, pp. 10597–10606. PMLR (2020)
24. Yu, J., Gao, H., Chen, Y., Zhou, D., Liu, J., Ju, Z.: Deep object detector with attentional spatiotemporal LSTM for space human-robot interaction. IEEE Trans. Hum.-Mach. Syst. **52**(4), 784–793 (2022)
25. Yu, J., Gao, H., Sun, J., Zhou, D., Ju, Z.: Spatial cognition-driven deep learning for car detection in unmanned aerial vehicle imagery. IEEE Trans. Cogn. Dev. Syst. **14**(4), 1574–1583 (2021)
26. Yu, J., Xu, Y., Chen, H., Ju, Z.: Versatile graph neural networks toward intuitive human activity understanding. IEEE Trans. Neural Netw. Learn. Syst. (2022)
27. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process. Lett. **23**(10), 1499–1503 (2016)