Research on improved MobileFaceNet facial recognition algorithm

Guangtai Zhang*, Shuliang Zhang, Xin He
School of Automation, Nanjing University of Science and Technology, Nanjing, Jiangsu, China

ABSTRACT

In this article, a lightweight face recognition algorithm is constructed, which is based on the improved MobileFaceNet. To improve recognition accuracy and meet real-time requirements under the premise of ensuring a lightweight model, the inverse residual network of the ECA-Net network and H-swish activation function is designed. ECA-Net network enhances network cross-channel learning ability to improve algorithm accuracy and replaces ECA-Net network activation function with H-swish to enhance model device applicability.

Keywords: MobileFaceNet, face recognition, lightweight model, identity authentication, unconstrained scenarios

1. INTRODUCTION

Face recognition, as an important biometric technology, finds extensive applications in the field of information security, particularly in the following areas: identification for documents and real-name magnetic card verification[1]; surveillance systems in customs, shopping malls, stations, airports, and banks[2,3]; public surveillance, enterprise, and residential security and management, such as facial access control attendance systems and facial recognition anti-theft doors; and matching of suspect photos[4,5].With the booming of deep convolutional neural networks in recent years, the performance of face detectors has been greatly improved[6].

There are many algorithm models for face recognition[7-10], including common networks such as DeepID and FaceNet. However, considering that facial recognition algorithms are limited by the deployment hardware resources and cannot meet the computational power requirements of these models[11,12], we have chosen lightweight facial recognition models suitable for system deployment, such as the MobileNet series[13,14], MobileFaceNet[15,16], and other lightweight models.

This article selects the MobileFaceNet model as the research object, and improves the model to address issues such as local occlusion and multi pose expressions in facial images, in order to improve the recognition performance of the model under unconstrained facial conditions.

2. ALGORITHM DESIGN BASED ON IMPROVED MOBILEFACENET

The MobileFaceNet[17] model is used for real-time facial recognition, based on the MobileNetV2 network framework. It uses globally separable convolution (GDConv) instead of average pooling layer operations and trains the network with arcface loss function to improve the facial recognition performance of the network model. The commonly used regression loss functions in object detectors are L1/L2 loss, smooth L1 loss, IoU loss and its variants[18].

During the experiment, it was found that MobileFaceNet had a decrease in facial recognition accuracy in unconstrained environments with side faces, local keypoint occlusion, and significant changes in facial expressions and postures. In order to improve the accuracy of the lightweight MobileFaceNet algorithm in this facial state, while considering the operational requirements of device performance and real-time performance during actual system operation, the MobileFaceNet network structure is improved by drawing on the advantages of lightweight MobileNetV3 network improvement.

2.1 ECA-Net

The ECA Net module appropriately increases cross-channel learning on the basis of SE Net learning channel weights. In order to improve performance without increasing network complexity, ECA Net adopts a local cross channel interaction method without dimensionality reduction. This method only designs a few parameters and can be effectively

*guangtai_zhang@njust.edu.cn

International Conference on Optics, Electronics, and Communication Engineering (OECE 2024), edited by Yang Yue, Proc. of SPIE Vol. 13395, 133951Y · © 2024 SPIE · 0277-786X · Published under a Creative Commons Attribution CC-BY 3.0 License · doi: 10.1117/12.3048257

implemented through one-dimensional convolution, avoiding the impact of SE Net dimensionality reduction on the learning channel, shown in Figure 1.

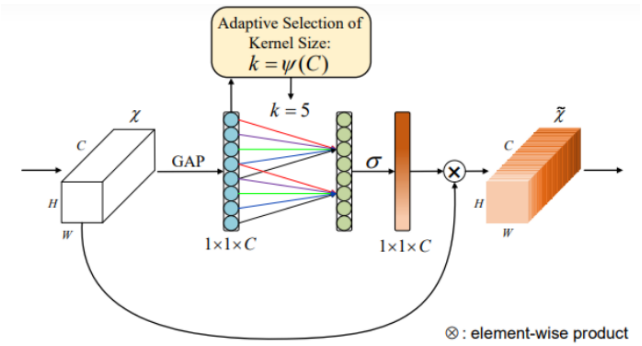

Figure 1. Structure of and ECA-Net modules.

The input features of the ECA Net network are pooled globally to obtain a 1×1×C dimensional feature vector, which is then weighted through a one-dimensional convolution operation of size K. K represents the number of cross channel interactions when the module calculates the weight of each channel, and the size of K is related to the number of input channels C, which is adaptively determined by the number of channels C. Through experiments, it was found that the optimal K value of the model is related to the depth of the network. The facial recognition model in this article is MobileFaceNet, which belongs to lightweight networks with shallow depth, so K is set to 3.

2.2 H-swish activation function

Convolutional neural networks commonly use the ReLU activation function to mitigate the vanishing gradient problem and expedite model convergence. ReLU sets all negative values to zero while leaving positive values unchanged. However, unbounded positive outputs in ReLU can be problematic. To address this, the ReLU6 activation function limits the maximum output value.

An alternative to ReLU is the swish non-linear activation function, known to enhance network accuracy. Swish is defined as:

$$swishx = x.\partial(x) \tag{1}$$

In MobileNetV3, the authors adopt the H-swish activation function to balance accuracy and computational efficiency, specifically within inverted residual structures. H-swish is defined as:

$$h - swish[x] = x\frac{Relu6(x+3)}{6} \tag{2}$$

Here, the function caps the value at 6 when x exceeds 6, optimizing accuracy and model suitability while preserving computational speed. Compared to ReLU6, H-swish incurs minimal additional computation, making it well-suited for mobile devices where efficiency is crucial. This strategic use of activation functions within MobileNetV3's architecture enhances network accuracy without excessively burdening computational resources.

2.3 Network structure design

Integrating the above ECA Net structure and H-swish activation function into the bottleneck structure of the MobileFaceNet facial recognition algorithm to improve the model's facial recognition performance in situations such as multi pose expression changes and local occlusion. Firstly, based on the reverse residual structure of the bottleneck network, an ECA Net module is added to improve the network channel feature extraction capability. At the same time, the H-swish activation function is introduced to replace the ReLU activation function in ECA Net to improve network accuracy and computational speed, while retaining the original network's use of the linear activation function to alleviate the problem of feature information loss caused by dimensionality reduction operations. The integration of ECA Net, H-swish activation function, and reverse residual structure is shown in Figure 2.

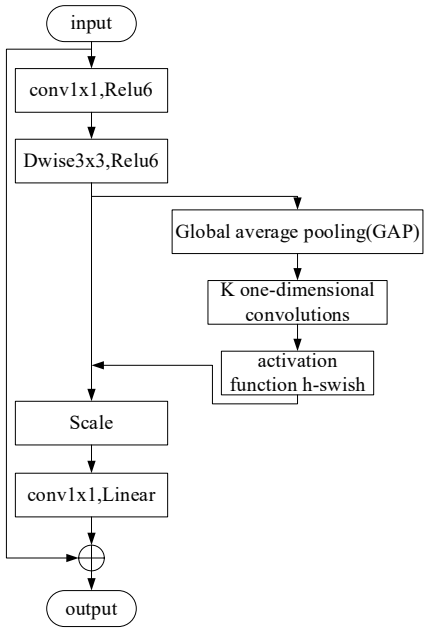

Figure 2. Schematic diagram of bottleneck module structure improvement.

To manage model complexity and computational demands, the ECA-Net structure is selectively integrated into the bottleneck layer. Additionally, the activation function within the ECA-Net is replaced with H-swish to strike a balance between accuracy and computational efficiency.

3. EXPERIMENTS AND ANALYSIS

In this section, we enhance the MobileFaceNet face recognition algorithm by drawing inspiration from the enhancements introduced in the MobileNetV3 network model. We incorporate the ECA-Net module to improve network accuracy and replace the activation function with H-swish to boost computational speed. Experimental comparisons between the original MobileFaceNet algorithm and our improved version are conducted to analyze and validate the effectiveness of these enhancements using experimental data.

3.1 Training dataset

The CASIA-WebFace dataset, sourced from the internet, comprises 494,414 facial images belonging to 10,575 individuals. This dataset is widely utilized in facial recognition algorithms due to its diverse range of conditions, including unconstrained settings, multiple poses, angles, and scenes for each individual's facial photos. Therefore, this study utilized the CASIA-WebFace dataset for network training.

However, since the dataset is collected through web scraping, it may contain low-quality images that do not meet face recognition standards, such as blurry images or those lacking visible faces.

To address this issue, the study employed preprocessing and cleaning techniques using the RetinaFace face detection algorithm to filter out images where no face could be detected. Organizing the dataset involved careful attention to image locations and corresponding facial annotation files, ensuring that images of the same individual were grouped in the same folder.

3.2 Model training

The face recognition model training in this study took place in the same device environment as the face detection model. However, there was a switch in the deep learning framework to PyTorch version 1.10.

The content of the training parameters for the facial recognition model is as follows: the training optimization method adopts Adam optimization method, the optimization method of random gradient descent uses momentum and adaptive learning rate to accelerate convergence speed, the momentum parameter used in the optimization method is set to 0.9, the

initial learning rate of the model is 0.001, the minimum learning rate is 0.0001, the learning rate attenuation method is cosine descent method, the weight attenuation is set to 0, the batch size is set to 16, the number of iterations is 100, and the training loss function is ArcFace.

3.3 Testing and analysis

To assess the improvement, experiments were conducted using MobileFaceNet as the backbone network, progressively integrating ECA-Net and transitioning the activation function to H-swish during model training. The weight file corresponding to the model with the lowest loss was retained for evaluation. Evaluations were performed on the LFW test dataset using these different model configurations.

The evaluation process began by loading the trained model and employing the RetinaFace face detection algorithm to determine face bounding box positions and 5 facial landmarks for face cropping. Utilizing annotation information from the LFW training dataset, the face recognition model extracted features for each pair of test images, resulting in face feature vectors. The Euclidean distance between the feature vectors of each test pair was calculated. A Euclidean distance threshold was set, and distances were iterated through. If the distance was below the threshold, the pair was classified as the same person; otherwise, they were classified as different. By comparing these classifications with ground truth labels, accuracy was computed for each threshold. The threshold yielding the highest accuracy was deemed optimal for the model.

The evaluation results, depicted in Figure 3, show that the original model achieved a test accuracy of 97.72% with an Euclidean distance threshold of 1.10. Models are mostly fast, but not accurate enough[19]. In contrast, the improved model achieved a test accuracy of 99.02% with a slightly lower Euclidean distance threshold of 1.08. These findings highlight the efficacy of the enhancements introduced to the MobileFaceNet model.

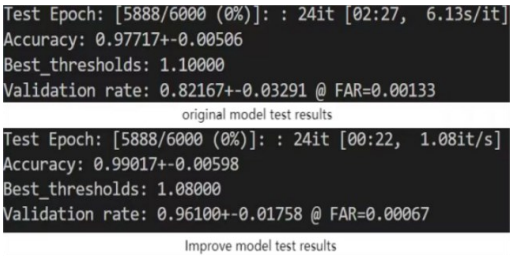

Figure 3. Test results of MobileFaceNet model and improved model.

Table 1 presents detailed test results for each experimental model, including detection accuracy (AP), model size, and inference speed (FPS). FPS values were determined by evaluating input images using a face detection and recognition model featuring ResNet50 as the backbone network. The test images used were identical to those employed for training the face detection model.

Table 1. Performance data comparison for models on the LFW dataset.

ECA-Net	H-swish	Average precision AP (%)
-	-	97.72
√	-	98.95
√	√	99.02

The comparative analysis of experimental data demonstrates that within the same experimental environment and parameter settings, introducing ECA-Net improves the accuracy of the original MobileFaceNet model on the LFW test set. Additionally, the H-swish activation function enhances the model's computational speed to a certain extent.

Regarding detection accuracy (AP), the original MobileFaceNet achieves 97.72% accuracy on the LFW test set. Introducing ECA-Net results in a 1.23% accuracy improvement, but it increases the model size from 4.7 MB to 4.85 MB. This expansion in network structure leads to larger model size, more parameters, and increased computational load, resulting in a slight decrease of 0.159 in inference speed (FPS). However, the use of the H-swish activation function improves model accuracy while ensuring that computational speed remains within an acceptable range.

4. CONCLUSIONS

This article mainly focuses on the research and experimental analysis of facial recognition algorithms. Taking the lightweight model MobileNet series as the research object, this paper analyzes the characteristics of separable convolution and reverse residual structures in network models, and combines the improvement characteristics of MobileNet series models to improve accuracy while ensuring network operation speed. In response to the multi pose and multi expression characteristics of faces in unconstrained scenes, the ECA Net structure is introduced in the reverse residual block to increase cross channel weight learning and improve network feature extraction ability. At the same time, the H-swish activation function is used to ensure model operation speed and improve accuracy. The CASIA WebFace dataset and LFW test dataset were used for data evaluation, and experimental data showed the effectiveness of the improved method.

REFERENCES

[1] Zhao, W., Chellappa, R., Phillips, P. J., et al., "Face recognition: A literature survey," ACM Computing Surveys, 35(4), 399-458 (2003).

[2] Hietmeyer, R., "Biometric identification promises fast and secure processing of airline passengers," ICAO Journal, 55(9), 10-11 (2000).

[3] Huang, T., Xiong Z. and Zhang, Z., [Face Recognition Applications], Handbook of Face Recognition, London: Springer, 617-638 (2011).

[4] Kc, G. S. and Karger, P. A., "Cryptology," ePrint Archive. Report, 2005.

[5] Hinton, G. E. and Salakhutdinov, R. R., "Reducing the dimensionality of data with neural networks," Science, 313, 504-507 (2006).

[6] Yu, Z., Huang, H., et al., "YOLO-FaceV2: A scale and occlusion aware face detector," Pattern Recognition, 155, (2024).

[7] Krizhevsky, A., Sutskever, I. and Hinton, G. E., "ImageNet classification with deep convolutional neural networks", NIPS, (2012).

[8] Long, J., Shelhamer, E. and Darrell, T., "Fully convolutional networks for semantic segmentation", CVPR, (2015).

[9] Ren, S., He, K., Girshick, R. and Sun, J., "Faster R-CNN: Towards real-time object detection with region proposal networks", NIPS, (2015).

[10] Toshev, A. and Szegedy, C., "DeepPose: Human pose estimation via deep neural networks", CVPR, (2014).

[11] Ioffe, S. and Szegedy, C., "Batch normalization: Accelerating deep network training by reducing internal covariate shift", ICML, (2015).

[12] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., et al., "Going deeper with convolutions", CVPR, (2015).

[13] Bell, S., Zitnick, C. L., Bala, K. and Girshick, R., "Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks", CVPR, (2016).

[14] Newell, A., Yang, K. and Deng, J., "Stacked hourglass networks for human pose estimation", ECCV, (2016).

[15] Jaderberg, M., Simonyan, K., Zisserman, A. and Kavukcuoglu, K., "Spatial transformer networks", NIPS, (2015).

[16] Chen, S., Liu, Y., Gao, X., et al., "Mobilefacenets: Efficient CNNS for accurate real-time face verification on mobile devices," Biometric Recognition: 13th Chinese Conference, CCBR 2018, Urumqi, China, August 11-12, Proceedings 13, Springer International Publishing, 428-438 (2018).

[17] Sandler, M., Howard, A., Zhu, M., et al., "Mobilenetv2: Inverted residuals and linear bottlenecks," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 4510-4520 (2018).

[18] Zhang, Y. F., Ren, W., Zhang, Z., et al., "Focal and efficient IOU loss for accurate bounding box regression," arXiv:2101.08158v2, (2021).

[19] Jin, H., Liao, S. and Shao, L., "Pixel-in-pixel net: Towards efficient facial landmark detection in the wild," International Journal of Computer Vision, 129(12), 1-21 (2021).