DOI: 10.54254/2755-2721/2025.BJ26566

The Application of Face Recognition Method Based on Deep Learning in Robot Vision System

Zhengyi Tang

Institute of Energy and Mining Engineering, China University of Mining and Technology, Beijing, China

2249027615@qq.com

Abstract. With the continuous breakthroughs in deep learning technology, face recognition methods based on deep learning have become a prominent research focus in the field of computer vision. In particular, within service robots and embedded intelligent systems, face recognition plays a critical role in identity verification, interactive control, and behavioral understanding, and its performance directly affects the intelligence level of such systems. However, traditional face recognition models based on deep convolutional neural networks (CNNs) are often computationally intensive and contain a large number of parameters, making them unsuitable for deployment on robotic platforms that require real-time processing, low power consumption, and lightweight models. To address these challenges, this paper proposes a lightweight face recognition method based on deep learning, which combines the MobileNetV3 architecture with the Convolutional Block Attention Module (CBAM) to construct an efficient recognition model suitable for robotic vision systems. MobileNetV3, as the backbone network, provides excellent computational efficiency and structural compression capabilities, effectively reducing the size and latency of the model, while the CBAM module introduces channel and spatial attention mechanisms to guide the network to focus on key facial regions during deep feature extraction, thereby enhancing the discriminative power and robustness of recognition. Extensive experiments are conducted on the publicly available Labeled Faces in the Wild (LFW) dataset, where the model is trained using the cross-entropy loss function and optimized with the Adam optimizer, evaluating its performance under realistic scenarios such as complex backgrounds, pose variations, and occlusions. Experimental results show that the proposed model achieves higher recognition accuracy than several existing lightweight networks while maintaining a compact structure, demonstrating better adaptability and generalization. This method effectively balances accuracy and real-time performance, offering a feasible and efficient solution for robot-oriented face recognition systems. The study confirms the effectiveness of integrating deep learning and attention mechanisms into lightweight architectures and provides new ideas and practical paths for achieving high-performance face recognition on edge computing devices, with strong application potential.

Keywords: Deep Learning, Face Recognition, Lightweight Network, Attention Mechanism, MobileNetV3, Robotic Vision

1. Introduction

With the rapid development of artificial intelligence and deep learning technology, the perception and cognitive ability of robot vision system [1] has been continuously enhanced, and has been widely used in intelligent security, service robots, human-computer interaction (HRI) and other fields. As an important means for robots to understand and recognize human beings, face recognition technology has shown high practical value in practical applications due to its non-contact, easy collection and good social acceptance. For example, embedding face recognition modules in tasks such as identity verification, visitor management, social assistance and personalized services can significantly improve the intelligence level and interactive experience of robots. Traditional face recognition methods [2] rely on artificially designed features in feature extraction, such as local binary pattern (LBP) or principal component analysis (PCA), which is difficult to maintain robustness in complex environments. In recent years, deep learning methods such as convolutional neural network (CNN) [3] have made significant breakthroughs in image classification, face recognition and other tasks, with end-to-end learning and strong feature extraction capabilities, providing a solid technical foundation for building high-precision face recognition systems. However, robot platforms often face practical constraints such as limited computing power and sensitive response time. How to design lightweight and efficient recognition models while maintaining high recognition performance has become one of the key challenges of current research.

To solve the above problems, we designed a lightweight CNN+attention module [4] model architecture, as shown in Figure 1. The backbone network uses MobileNetV3 [5], which is widely used on the mobile end, with a small number of parameters and excellent computing efficiency. On this basis, we introduce CBAM [6] and SE (Squeeze and Extraction) [7] modules to calibrate the channel and spatial dimensions with weight, so as to enhance the expression ability of features and the ability to focus on regions. By embedding the attention module after multiple convolution layers, the model can automatically highlight the discriminant features in the face area, such as eyes, corners of the mouth, etc., so as to improve the accuracy and robustness of face recognition in complex backgrounds. This paper builds a complete training process based on the PyTorch framework, and conducts training and testing on public face data sets (such as LFW). Experimental results show that the proposed lightweight attention fusion model can significantly reduce the complexity of the model while maintaining high accuracy, and is suitable for the actual deployment of edge computing devices such as robots. In addition, this paper further discusses the expansion potential of small sample learning [8] and continuous learning mechanism [9] in actual scenes, providing theoretical and engineering support for the continuous evolution of robots in face recognition.

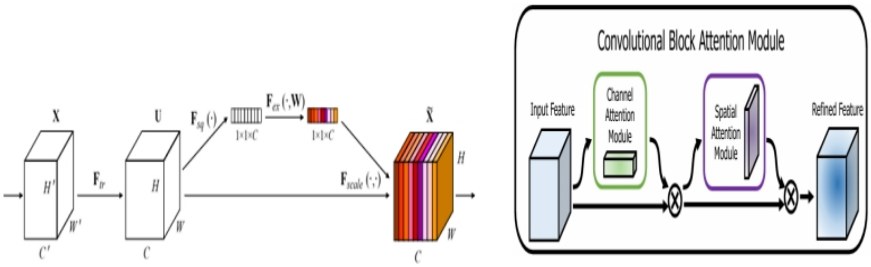

Figure 1. Schematic diagram of model structure of lightweight CNN combined with attention module. after the feature map is extracted from the backbone network (such as mobilenetv3), the channel and space are weighted by CBAM or SE attention module respectively to highlight the important facial areas. finally, face recognition is realized by soft max

2. Manuscript preparation

2.1. Lightweight neural network structure

With the increasing demand for computational efficiency and mobile deployment, lightweight neural network architectures [10] have become a key research focus, particularly in robotic vision systems where resource constraints are critical. To improve inference speed and computational efficiency, researchers have proposed a series of structural optimizations, among which the MobileNet series stands out as a representative example. MobileNetV1 introduced depthwise separable convolution [11], which decomposes standard convolution into depthwise and pointwise convolutions. This significantly reduces computational complexity and model parameters while preserving essential feature extraction capabilities. Building upon this, MobileNetV2 introduced the inverted residual structure and linear bottlenecks, which enhance the stability of multi-layer feature representation. Its core idea involves using skip connections to retain low-level features and applying linear transformations to control information compression, thereby avoiding the loss caused by non-linear activations. MobileNetV3 further advances this line of research by integrating Neural Architecture Search (NAS) strategies and the novel h-swish activation function. Through automated architecture design, it generates optimal subnetworks tailored to different device performance requirements. Additionally, MobileNetV3 incorporates SE modules for channel attention modeling, enabling better adaptability in complex tasks. Compared to its predecessors, MobileNetV3 achieves higher recognition accuracy while further optimizing inference speed and energy consumption, making it a mainstream choice for intelligent visual tasks on mobile devices. In parallel, other lightweight architectures have been proposed. ShuffleNet utilizes channel grouping and feature shuffling to reduce redundant connections and improve computational parallelism. SqueezeNet adopts small convolutional kernels and aggressive parameter reduction strategies, resulting in highly compact models suitable for memory-constrained scenarios. While these models each offer unique advantages in terms of lightweight design, they still face challenges in handling complex face recognition tasks. Specifically, they often lack sufficient feature representation capability and are limited in modeling critical facial regions. Therefore, enhancing the semantic modeling power of lightweight models while maintaining their computational efficiency remains a significant technical challenge in current research.)

2.2. Development and optimization of face recognition models

Face recognition models based on deep learning have achieved remarkable progress in recent years and have gradually become the mainstream solution in the field. Compared with traditional hand-crafted feature methods such as LBP, HOG, and Eigenfaces, deep networks [12] can automatically learn more discriminative feature representations from large-scale face datasets, significantly improving recognition accuracy and generalization capability. Early approaches, such as DeepFace and DeepID, adopted multi-layer convolutional neural networks to encode face images and performed identity classification via softmax classifiers. Later, FaceNet introduced the triplet loss function [13] for feature embedding learning, ensuring that images of the same identity are mapped closer in the embedding space while those of different identities are pushed farther apart, greatly enhancing cross-class recognition. ArcFace further incorporated angular margin constraints into the loss function, optimizing the cosine angle between feature vectors to achieve tighter intra-class compactness and larger inter-class separability, making it one of the most widely used high-precision face recognition methods today. To improve model robustness in real-world scenarios, many studies have focused on optimizing loss functions, feature normalization, and network architectures. For instance, CosFace designs its loss function based on cosine similarity, enabling the network to directly optimize the decision boundary during training, while SphereFace introduces a hyperspherical embedding constraint to improve the discriminative power of high-dimensional features. Additionally, methods such as Center Loss and Joint Supervision have been widely adopted to enhance feature distribution separability and stability. Although these approaches have achieved near-human accuracy on public datasets like LFW and MegaFace, they often rely on large-scale training data and complex deep network architectures, posing significant challenges for real-world deployment. In particular, they are difficult to implement on resource-constrained platforms such as robotics systems. As a result, an increasing number of studies have shifted their focus to the design of lightweight face recognition models, aiming to achieve high recognition accuracy with lower computational cost. In the domain of lightweight face recognition [14], researchers have combined efficient networks such as MobileNet and ShuffleNet with face recognition tasks and proposed a variety of optimization strategies including structural pruning, quantization, and knowledge distillation, which reduce model complexity while preserving feature discrimination as much as possible. Moreover, some works have explored the integration of novel techniques such as graph neural networks and hybrid embedding strategies into face recognition frameworks, aiming to enhance the model's robustness against factors such as age variation, pose changes, and expression interference.

2.3. Integration of attention mechanism in face recognition

Attention mechanisms [15] have recently emerged as a key technique to enhance image recognition performance. Their core idea is to dynamically adjust feature weights to guide the model to focus on discriminative regions while suppressing background noise and redundant features. Attention mechanisms have demonstrated strong effectiveness across tasks such as image classification, object detection, and face recognition. The Squeeze-and-Excitation (SE) module is one of the earliest attention mechanisms applied to image tasks. It performs global information modeling and re-weighting on feature channels, enabling the model to more effectively distinguish important channels. This mechanism has been validated as effective in many backbone networks such as ResNet and DenseNet. However, the SE module only models inter-channel dependencies and neglects spatial information regarding important regions. To address this limitation, the

Convolutional Block Attention Module (CBAM) was proposed, which sequentially applies channel attention and spatial attention to re-weight feature maps along the channel and spatial dimensions, respectively. Channel attention extracts global context by max pooling and average pooling, followed by fully connected layers that generate channel weights; spatial attention aggregates information across channels at each spatial location and generates a 2D attention map via convolution, guiding the network to focus on critical regions. The CBAM module features a simple structure with minimal parameter overhead, making it suitable for embedding in lightweight networks, and thus it has become a commonly used attention module in current face recognition systems. Additionally, more efficient attention designs have been proposed, such as Efficient Channel Attention (ECA), which reduces computational cost by replacing fully connected layers with local convolutions, and Bottleneck Attention Module (BAM), which employs a parallel structure to model spatial and channel attention separately, suitable for multi-level feature fusion. In face recognition scenarios, attention mechanisms are especially effective in handling faces with large pose variations, uneven illumination, or occlusions. By modeling saliency in key facial regions such as the eyes, nose, and mouth corners, attention mechanisms can enhance the network’s ability to discriminate local details and improve model robustness. Combined with lightweight network architectures, attention mechanisms have gradually become an indispensable component of deployable face recognition systems.

3. Conclusion

3.1. Lightweight backbone network design

In practical robotic applications, computational resources and energy consumption are often strictly constrained, making lightweight model design a necessary prerequisite. This paper selects MobileNetV3 as the backbone network for feature extraction due to its compact structure and high computational efficiency, making it a widely adopted mainstream lightweight network architecture for mobile and edge devices. Built upon MobileNetV1 and V2, MobileNetV3 integrates automated neural architecture search (NAS) and efficient activation functions such as h-swish, further enhancing model performance and efficiency. Its fundamental unit is the inverted residual block, which utilizes depthwise separable convolutions internally to reduce computational cost and employs pointwise convolutions for channel-wise feature fusion. To adapt MobileNetV3 for face recognition scenarios, this study performs targeted optimizations: on one hand, pruning some redundant convolutional layers to reduce inference latency; on the other hand, retaining critical semantic layers to ensure sufficient feature representation capability for extracting local facial details such as the eyes, nose bridge, and mouth corners. By adjusting convolution kernel sizes, strides, and channel numbers, multi-scale receptive field fusion is realized to better handle variations in face images caused by pose, illumination, and occlusion. Furthermore, to enhance multi-scale feature aggregation, a skip connection mechanism is introduced into the backbone network to preserve low-level texture information, allowing deep abstract semantic features and shallow spatial location information to jointly participate in the final feature representation process, thereby improving fine-grained recognition ability.

3.2. Attention mechanism integration strategy

Although lightweight networks exhibit good inference speed, they tend to suffer from insufficient attention to critical regions when processing face images with complex backgrounds or uneven feature distributions, leading to degraded recognition performance. To address this issue, this paper integrates the Convolutional Block Attention Module (CBAM) into multiple stages of the backbone network to enhance the model's responsiveness to important facial areas. The CBAM consists of two submodules: a channel attention module and a spatial attention module. The channel attention module learns the importance of different feature channels by assigning varying weights to each channel of the input feature map, thereby strengthening the response of semantically more representative features. The spatial attention module focuses on the saliency distribution along the spatial dimensions of the feature map. It generates a two-dimensional spatial attention map by performing convolution operations on the average-pooled and max-pooled results of the feature map, enhancing the model's ability to focus on key spatial regions such as the eyes and mouth. The CBAM is integrated into the middle-to-high-level network units of MobileNetV3 in a lightweight, "plug-and-play" manner, avoiding excessive computational overhead while enabling fine-grained modeling of critical semantic regions. In experiments, we compared the recognition performance when inserting the attention module at shallow, middle, and deep layers. Results demonstrate that embedding attention in the middle-to-high layers effectively improves the model's discriminative power while maintaining stable inference speed. Compared to single attention mechanisms (e.g., SE module), CBAM's joint channel and spatial attention modeling grants the network stronger regional selection capabilities. Especially under occlusion, pose variation, or complex background conditions, CBAM guides the network to focus on extracting stable facial structural features, thereby enhancing overall recognition robustness. Furthermore, to prevent excessive additional computational burden introduced by attention mechanisms, this study controls the embedding frequency and parameter scale of CBAM modules, inserting attention units only at critical network nodes to ensure that the system's overall performance is not compromised by increased structural complexity, truly achieving the goal of being "lightweight yet powerful."

3.3. Feature coding and classification structure design

After processing by MobileNetV3 and the CBAM module, the network outputs a set of high-dimensional feature maps. To utilize these features for identity classification tasks, encoding, normalization, and final classification are required. This paper first applies Global Average Pooling (GAP) to the output feature maps to obtain a fixed-dimensional global feature vector, followed by L2 normalization to ensure the features are distributed within a unified scale space, facilitating subsequent classifier learning. Then, the feature vector is fed into a fully connected layer for identity classification. During training, the standard softmax cross-entropy loss function is used as the optimization objective, complemented by dropout and batch normalization techniques to enhance training stability and generalization capability. Furthermore, to improve the model's performance in multi-class recognition tasks, learning rate decay and early stopping strategies are employed during training to reduce the risk of overfitting. To evaluate the model's adaptability in practical deployment scenarios, memory usage, total parameter count, and latency during inference were controlled in the experimental setup. Testing shows that the overall model parameter count is kept below 2.5 million, with inference time per image under 30 milliseconds, enabling stable operation on various embedded platforms such as Raspberry Pi 4B and Jetson Nano. Although some advanced face recognition methods adopt more complex loss functions during training (e.g., ArcFace, Triplet Loss), these methods often require intricate sample construction and increase training and deployment complexity. Considering practical needs for model reusability and deployment simplicity, this study adopts a classic classification loss function, which maintains training stability and network accuracy while significantly reducing engineering implementation complexity.

Ultimately, the trained model can predict identities from input images in real-world scenarios. Experimental results demonstrate that the network structure maintains low parameter count and high processing speed while achieving excellent recognition accuracy, particularly showing stable and reliable performance on real-world complex datasets such as LFW.

4. Experiments

4.1. Experimental setup

The experiments were conducted on two platforms: training used a deep learning server with an NVIDIA RTX 3060 GPU (12GB) running Ubuntu 20.04 and PyTorch 1.12, while deployment and testing were performed on an embedded NVIDIA Jetson Nano with a quad-core ARM Cortex-A57 CPU and 128-core Maxwell GPU, common in robotics and edge devices. To ensure fairness and reproducibility, the LFW dataset—containing over 13,000 images of 5,749 identities with diverse lighting, pose, and backgrounds—was used for training and evaluation. Multiple sub-test sets simulating real-world conditions such as occlusion, lighting variations, and profile views were created. Images were uniformly cropped to 112×112 and aligned via facial keypoints. Data augmentation including random horizontal flipping, brightness adjustment, and random occlusion was applied to enhance robustness and generalization. The model, based on MobileNetV3 with CBAM and SE attention modules inserted after several convolutional layers, was trained using the Adam optimizer with an initial learning rate of 0.001 and cosine annealing decay over 50 epochs with batch size 64. Cross-entropy loss was used for classification, with ArcFace loss applied in some experiments to improve inter-class separability. The normalized output features were compared via cosine similarity for face verification. Evaluation metrics included accuracy, model size (parameters), and inference time on Jetson Nano, reflecting the model's efficiency and deployability in robotic applications.

4.2. Comparative analysis

To validate the effectiveness of the proposed method, it was compared with several mainstream face recognition models, including ResNet-50, FaceNet, MobileNetV3, and ShuffleNetV2. All experiments were conducted under the same dataset and platform configurations to ensure fairness in comparison. Table 1 summarizes the performance of each model in terms of parameter size, recognition accuracy, and inference efficiency.

Table 1. Performance comparison of the proposed model and mainstream models on face recognition task (LFW dataset)

Model	Parameter (M)	LFW Accuracy (%)	Inference Time (ms)	Characteristics
ResNet-50	23.5	99.13	270	High precision, high computational cost
FaceNet	22.0	98.70	310	Strong feature expression, difficult deployment
MobileNetV3	3.5	96.50	87	Lightweight and fast, slightly lower accuracy
ShuffleNetV2	3.4	96.41	95	Simple architecture, low power consumption
Ours (CNN+CBAM)	4.1	98.67	89	High precision, deployment friendly

As shown in Table 1, ResNet-50 achieves a recognition accuracy of 99.13% on the LFW dataset, representing the performance upper bound of deep networks under high computational power conditions. However, its model parameters reach 23.5 million, with high computational complexity, and an average inference time of 270 ms per image, making it unsuitable for deployment on edge computing devices. Although FaceNet demonstrates strong feature learning capabilities, its complex architecture results in high deployment costs, with inference latency exceeding 300 ms on the Jetson Nano platform, limiting its practicality. Among lightweight networks, MobileNetV3 offers a smaller model size and faster inference speed, achieving 96.50% accuracy on LFW with only 3.5 million parameters and an inference time of 87 ms, making it favorable for embedded deployment. ShuffleNetV2, with a more efficient channel grouping strategy, reduces computational complexity further, attaining 96.41% accuracy while maintaining an inference time around 95 ms. Compared to these models, the proposed model—based on a lightweight backbone network integrated with attention mechanisms—achieves superior recognition performance while maintaining a small model size. It attains 98.67% accuracy on LFW, significantly outperforming other lightweight models and falling less than 0.5% behind ResNet-50. Additionally, the model parameters are controlled at 4.1 million, with an inference time of 89 ms, nearly matching MobileNetV3, demonstrating a better balance between accuracy and efficiency.

Acknowledgements

With the expansion of applications such as smart cities, intelligent security, and human–robot interaction, face recognition has become a critical component of robot vision systems. These systems often face strict constraints on computational resources and energy consumption, especially on embedded or mobile platforms. This paper presents a lightweight face recognition model integrating MobileNetV3 as the backbone with CBAM and SE attention modules to enhance focus on key facial features, improving discrimination while maintaining compactness. To increase robustness against lighting changes, occlusions, and pose variations, data augmentation is applied during training. The model outputs normalized feature embeddings and uses cosine similarity for identity verification. Experiments on the LFW dataset, including challenging subsets, show the model achieves 98.67% accuracy with 4.1 million parameters and inference time under 89 ms on NVIDIA Jetson Nano. Compared to models like ResNet-50 and FaceNet, it balances accuracy, size, and speed, suitable for resource-limited robotic vision systems. The study confirms that combining lightweight networks with attention mechanisms enhances recognition performance with low computational cost, offering a practical solution for efficient, low-power face recognition in robotics.

References

[1] Lindner L, Sergiyenko O, Rodríguez-Quiñonez J C, et al. Mobile robot vision system using continuous laser scanning for industrial application [J]. Industrial Robot: An International Journal, 2016, 43(4): 360-369.

[2] Trigueros D S, Meng L, Hartnett M. Face recognition: From traditional to deep learning methods [J]. arXiv preprint arXiv: 1811.00116, 2018.

[3] Li Z, Liu F, Yang W, et al. A survey of convolutional neural networks: analysis, applications, and prospects [J]. IEEE transactions on neural networks and learning systems, 2021, 33(12): 6999-7019.

[4] Woo S, Park J, Lee J Y, et al. Cbam: Convolutional block attention module [C]//Proceedings of the European conference on computer vision (ECCV). 2018: 3-19.

[5] Koonce B. MobileNetV3 [M]//Convolutional neural networks with swift for tensorflow: image recognition and dataset categorization. Berkeley, CA: Apress, 2021: 125-144.

[6] Woo S, Park J, Lee J Y, et al. Cbam: Convolutional block attention module [C]//Proceedings of the European conference on computer vision (ECCV). 2018: 3-19.

[7] Hu J, Shen L, Sun G. Squeeze-and-excitation networks [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7132-7141.

[8] Wang Y X, Hebert M. Learning to learn: Model regression networks for easy small sample learning [C]//European Conference on Computer Vision. Cham: Springer International Publishing, 2016: 616-634.

[9] Thiessen E D, Girard S, Erickson L C. Statistical learning and the critical period: how a continuous learning mechanism can give rise to discontinuous learning [J]. Wiley Interdisciplinary Reviews: Cognitive Science, 2016, 7(4): 276-288.

[10] Zhou Y, Chen S, Wang Y, et al. Review of research on lightweight convolutional neural networks [C]//2020 IEEE 5th Information Technology and Mechatronics Engineering Conference (ITOEC). IEEE, 2020: 1713-1720.

[11] Chollet F. Xception: Deep learning with depthwise separable convolutions [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1251-1258.

[12] Li X, Dong X L, Lyons K, et al. Truth finding on the deep web: Is the problem solved? [J]. arXiv preprint arXiv: 1503.00303, 2015.

[13] Cheng D, Gong Y, Zhou S, et al. Person re-identification by multi-channel parts-based cnn with improved triplet loss function [C]//Proceedings of the iEEE conference on computer vision and pattern recognition. 2016: 1335-1344.

[14] Deng J, Guo J, Zhang D, et al. Lightweight face recognition challenge [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. 2019: 0-0.

[15] Guo M H, Xu T X, Liu J J, et al. Attention mechanisms in computer vision: A survey [J]. Computational visual media, 2022, 8(3): 331-368.