Online-Journals.org

iJIM International Journal of Interactive Mobile Technologies

iJIM | eISSN: 1865-7923 | Vol. 17 No. 23 (2023) | OPEN ACCESS

https://doi.org/10.3991/ijim.v17i23.40867

PAPER

A Review of Deep Convolutional Neural Networks in Mobile Face Recognition

Jing Chi[1], Chin Kim On[1](✉), Haopeng Zhang[2], Soo See Chai[3]

[1]Faculty of Computing and Informatics, University Malaysia Sabah, Sabah, Malaysia

[2]School Information and Electrical Engineering, Hebei University of Engineering, Hebei, China

[3]Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak, Sarawak, Malaysia

kimonchin@ums.edu.my

ABSTRACT

With the emergence of deep learning, Convolutional Neural Network (CNN) models have been proposed to advance the progress of various applications, including face recognition, object detection, pattern recognition, and number plate recognition. The utilization of CNNs in these areas has considerably improved security and surveillance capabilities by providing automated recognition solutions, such as traffic surveillance, access control devices, biometric security systems, and attendance systems. However, there is still room for improvement in this field. This paper discusses several classic CNN models, such as LeNet-5, AlexNet, VGGNet, GoogLeNet, and ResNet, as well as lightweight models for mobile-based applications, such as MobileNet, ShuffleNet, and EfficientNet. Additionally, deep CNN-based face recognition models, such as DeepFace, DeepID, FaceNet, and SphereFace, are explored, along with their architectural characteristics, advantages, disadvantages, and recognition accuracy. The results indicate that many scholars are researching lightweight face recognition, but applying it to mobile devices is impractical due to high computational costs. Furthermore, noise label learning is not robust in actual scenarios, and unlabeled face learning is expensive in manual labeling. Finally, this paper concludes with a discussion of the current problems faced by face recognition technology and its potential future directions for development.

KEYWORDS

computer vision, deep learning, deep convolutional neural network, face recognition, lightweight deep CNN models

1 INTRODUCTION

Face recognition is a biometric identification technology that utilizes physiological features to extract human facial features via a computer and authenticate identity based on these features [1]. With the continuous development of deep learning technology, especially Deep Convolutional Neural Network (DCNN), the field of face recognition has seen significant progress in recent years. As a non-contact and easy-to-implement technology, face recognition is not only applicable to mobile

Chi, J., Kim On, C., Zhang, H., Chai, S.S. (2023). A Review of Deep Convolutional Neural Networks in Mobile Face Recognition. *International Journal of Interactive Mobile Technologies (iJIM)*, 17(23), pp. 4–19. https://doi.org/10.3991/ijim.v17i23.40867

Article submitted 2023-04-26. Revision uploaded 2023-07-24. Final acceptance 2023-10-21.

devices, but it can also effectively mitigate the risk of contact transmission in the current context of pandemic prevention and control. This technology plays a critical role in enhancing security and surveillance capabilities by providing automated recognition solutions such as traffic surveillance, access control devices, biometric security systems, and attendance systems [2].

Face recognition involves three general steps: face detection, feature extraction, and face classification [3]. Face detection extracts information about facial position, size, and key point position from face images. Feature extraction uses a Convolutional Neural Network (CNN) to extract facial features. Face classification performs discriminative classification by comparing the extracted facial features with the feature information in the database [4]. Although the use of CNN has been proven effective in face recognition, there is still room for improvement. The performance of these models depends on the variety of algorithms used. This article focuses on describing and comparing commonly used CNN algorithms in face recognition, including their architectural characteristics, advantages, disadvantages, and recognition accuracy.

The remainder of this paper is organized as follows: In Section 2, we provide a brief overview of classic CNN models, including their architectural characteristics, advantages, and disadvantages. We also summarize their error rates in classification tasks. In Section 3, we focus on analyzing face recognition models and their architectural characteristics, advantages, disadvantages, and recognition rates in face datasets. Finally, we discuss the challenges faced by current face recognition technology and propose future development directions for face recognition.

2 DEEP CNN

In recent years, the development of deep learning has led to a transformation in the theoretical models used for face recognition. Traditional artificial feature extraction models have been replaced by CNN models with higher quality and accuracy [5] [6]. Face recognition is achieved by feeding a face image into the CNN model, which extracts multi-dimensional features that are then compared to the facial features stored in the database. The most commonly used DCNN models include LeNet-5 [7], AlexNet [8], VGGNet [9], GoogleNet [10], and ResNet [11]. In addition, there are also lightweight CNNs, such as MobileNet [12][13][14], ShuffleNet [15][16], and EfficientNet [17][18], which will be briefly explained in the following subsections.

2.1 LeNet-5

The LeNet-5 model, proposed in 1998, is one of the earliest CNN models, comprising of three convolutional layers, two pooling layers, and two fully connected layers, using the Sigmoid activation function [7]. Thanks to its innovative architecture, LeNet-5 played a crucial role in advancing deep learning. Its multiple layers of convolution and pooling operations paved the way for the development of more complex and powerful CNN models. Additionally, LeNet-5 was designed specifically for handwritten digit recognition, a challenging task at the time, and achieved high accuracy on this task. Its success demonstrated the potential of CNNs for solving complex pattern recognition problems and inspired further research in the field. However, technological limitations, particularly hardware processing speed and stability, hindered the widespread adoption of CNN models, making it difficult for researchers to extend and apply these models further [19].

2.2 AlexNet

The AlexNet model won the first prize in the 2012 ImageNet Large Scale Visual Recognition Challenge (ILSVRC) and greatly advanced the development of deep learning [20]. It has five notable characteristics. Firstly, the ReLU activation function was introduced to address the problem of gradient disappearance or explosion, which significantly speeds up network training compared to the previously used Sigmoid activation function. Secondly, the Dropout regularization technique was proposed to alleviate the model's overfitting problem by randomly deactivating some neurons during data training, reducing the number of network parameters. This approach enables all neurons to learn more comprehensive features, improving the network's generalization ability [21]. Thirdly, CUDA was used to leverage the GPU's powerful parallel computing ability to accelerate the Neural Network (NN) training. Fourthly, data augmentation techniques were applied to prevent overfitting and encourage the NN to recall essential features while also generating additional data. Lastly, the AlexNet model uses overlapping max pooling to reduce the spatial size of the output volume. However, the AlexNet model has some limitations. One limitation is that it requires a large amount of labeled data for training, which can be time-consuming and expensive to obtain. Additionally, the model is computationally expensive, and it may not be suitable for use on devices with limited processing power. Another limitation is that the model may not generalize well to new or unseen data, as it can be prone to overfitting. Finally, the AlexNet model has a large number of parameters, which can make it difficult to train and optimize.

2.3 VGGNet

In 2014, the Visual Geometry Group from the University of Oxford proposed the VGG network. VGG16 to VGG19 are variants of the network that have been widely adopted in computer vision tasks. Compared to AlexNet, VGG utilizes smaller convolution kernels to deepen the network, which improves recognition accuracy and enhances the generalization ability of CNNs. However, smaller convolution kernels require the computation of a substantial number of parameters at each layer, which slows down network learning and relies on the powerful computing ability of GPUs. The input to VGG is an RGB image of size 224 × 224 × 3. Initially, the average value is calculated for all images in the training set, and the normalized images are input into VGG. Then, the images are convolved with 3 × 3 or 1 × 1 kernels. Finally, classification is performed through three fully-connected layers. VGG has diverse architectures with layer numbers ranging from 11 to 19. As the number of layers increases, the accuracy of VGG reaches a bottleneck at the 16th layer, which tends to saturate [22].

2.4 GoogLeNet

The Google research team developed GoogLeNet, which won the 2014 ILSVRC and improved network performance by deepening the number of layers further. The GoogLeNet network introduces the inception module, which is used in nine inception modules, resulting in a total of 18 layers. The inception module first convolves the input with a 1 × 1 kernel and a 3 × 3 window for maximum pooling. Next, separate convolutions are performed with 1 × 1, 3 × 3, and 5 × 5 kernels, and the resulting feature maps of different scales are fused. Finally, the two convolution results are combined to obtain the fused result. GoogLeNet has further deepened the network compared to previous networks, enabling the number of layers to reach 22. However, GoogLeNet's parameter quantity is only 1/12 of that of AlexNet while being approximately 10% more accurate than AlexNet in ImageNet, as reported in [23][24][25].

2.5 ResNet

The ResNet model was proposed in 2015 by [11]. As the depth of a network increases, the gradient can either vanish or explode in a network composed of stacked layers, making the model difficult to train [26]. In face recognition, it has been observed that the 56-layer network model has a higher error rate than the 20-layer model on both the training and test sets, indicating that adding more layers degrades the network performance. The introduction of ResNet helps alleviate the problem of training very deep networks. Skip connections are used in ResNet to transmit data to deeper NN levels to alleviate performance degradation. This ensures the integrity of information, improves training speed and efficiency, without increasing the computational burden of the network parameters [27].

2.6 MobileNet

The Google research team developed a model called MobileNet, a lightweight network model for better application of network models on mobile terminals and embedded devices. It mainly utilizes depthwise separable convolution, which splits the standard convolution into depthwise convolution and pointwise convolution. With depthwise convolution, different convolution kernels are used for each channel, and one input channel corresponds to one convolution kernel. With pointwise convolution, 1 × 1 kernels perform standard convolution operations. The depthwise separable convolution strategy greatly reduces network model parameters and computational complexity, albeit with a slight accuracy loss [28].

MobileNetv2, an extension of MobileNet, incorporates Inverted Residuals and Linear Bottlenecks borrowed from the ResNet architecture. The Inverted Residuals increase dimensionality via a 1 × 1 convolution kernel and use the ReLU6 nonlinear activation function, which can be more robust in low-precision computations. After a 3 × 3 convolution kernel, dimensionality reduction is performed using a 1 × 1 convolution kernel. Since ReLU loses extensive information during high-to-low dimensionality transformation [29][30], a linear activation function is used in MobileNetv2. Compared to MobileNet, MobileNetv2 features a smaller model, faster detection, and higher accuracy.

MobileNetv3 combines MobileNet's depthwise separable convolution and MobileNetv2's inverted residual architecture with linear bottlenecks and is also integrated with a lightweight attention module based on the Squeeze and Excitation [31] architecture. MobileNetv3 can automatically assign weights during training, thereby increasing the weights of valid feature maps while suppressing the weights of useless feature maps. The paper also proposes to search for the global network architecture through platform-aware neural architecture search (NAS)-based optimization of each network block and to search for the number of filters by exploiting the NetAdapt algorithm to ensure the optimal model can be found effectively. MobileNetv3 also introduces H-swish, a new nonlinear activation function. The H-swish function improves swish to reduce the computational overhead, which can effectively improve network accuracy [32].

2.7 ShuffleNet

ShuffleNet is a lightweight network model developed by Megvii's technology team in 2018. It utilizes two core operations: channel shuffle and pointwise group convolutions. The channel shuffle operation involves two stacked group convolutions. The group convolution channels do not communicate with each other, which can compromise the network's feature extraction ability. However, by using channel shuffle, ShuffleNet can effectively overcome this limitation. Then, the feature maps of each channel are allocated in an orderly manner, which fully utilizes the advantages of group convolution. Experimental results have demonstrated that in ImageNet classification, ShuffleNet outperforms MobileNet in both accuracy and computational complexity [15].

In ShuffleNetv2, the authors conducted four experiments and derived four guidelines: (1) make the numbers of input and output channels of the convolution equal; (2) consider that group convolution increases memory access cost (MAC); (3) avoid using a fragmented network architecture; and (4) reduce element-wise operations. The authors have successfully improved the ShuffleNet model and proposed ShuffleNetv2 by following these guidelines. ShuffleNetv2 introduces a Channel Split operation, and the basic units of ShuffleNet.

2.8 EfficientNet

In 2019, a team proposed the EfficientNet model [17] which improved accuracy by balancing resolution, depth, and width dimensions while saving computing resources. The model consists of multiple MBConv (mobile inverted bottleneck convolution) blocks, which have a structure similar to MobileNetv2's Inverted Residuals structure. The MBConv structure begins with a 1×1 pointwise convolution followed by a k×k depthwise convolution. Next, the Squeeze and Excitation attention selectively amplifies valuable feature channels and suppresses useless ones before a final 1×1 pointwise convolution is used to ensure the channel dimension is equal to the input channel dimension. Batch Normalization is applied after each convolution operation, and the Swish activation function is used. However, training EfficientNet becomes slow when increasing the resolution of training images or using Depthwise convolutions in shallow layers. In response, [18] presented EfficientNetv2 at CVPR in 2021, which proposes an improved progressive learning method that adjusts the canonical method dynamically based on the size of the training image. The studies reveal that using this approach has enhanced both training speed and accuracy. EfficientNetv2 replaces the shallow MBConv structure with the Fused-MBConv structure. Fused-MBConv uses standard convolution of 3×3 instead of depthwise convolution.

2.9 Summary of CNN models

The earliest CNN model, LeNet-5, had a simple architecture but suffered from overfitting. To address this issue, AlexNet incorporated dropout, which helped accelerate network training and convergence, marking a significant milestone for deep learning. VGG, on the other hand, increased model depth and accuracy by utilizing small convolution kernels, but at the expense of greater computational cost. The GoogLeNet improved accuracy by incorporating the Inception architecture, which reduces the number of parameters in the network. ResNet addressed the vanishing and exploding gradient problem in deep networks with its residual module. The lightweight MobileNet model utilized depthwise separable convolution technology to reduce parameters at a slight accuracy loss. MobileNetv2 introduced inverted residuals and linear bottlenecks, and MobileNetv3 used an attention model to improve accuracy and reduce parameters. ShuffleNetv1 introduced channel communication technology and group convolution modules to resolve non-communication issues between channels, which were further improved upon in ShuffleNetv2 with the addition of four rules for creating more effective networks. The EfficientNet model improved accuracy by balancing the resolution, depth, and width dimensions, and its successor, EfficientNetv2, improved training speed and performance further. Table 1 shows the Top-1 error rate on the validation set and the Top-5 error rate on the test set for different models in ImageNet classification, where "–" indicates that there is no corresponding experimental result in the corresponding reference.

Table 1. Error rates of different models in ImageNet

Model	Number of Parameters (M)	Top-1 Error Rates (Test, %)	Top-5 Error Rates (Test, %)
LeNet-5	0.06	–	–
AlexNet	60	36.7	15.3
VGG	138	–	7.3
GoogLeNet	5	–	6.67
ResNet-152	117	19.38	3.57
MobileNet v1-224	4.2	29.4	–
MobileNet v2(1.4)	3.4	25.3	–
MobileNet v3(1.0)	5.4	24.8	–
ShuffleNet v1 2x	5.4	26.3	–
ShuffleNet v2 2x	7.4	25.1	–
EfficientNet-B0	5.3	22.9	6.7
EfficientNetv2-S	22	16.1	–

3 FACE RECOGNITION BASED ON DCNN

The use of DCNN in computer vision and image recognition has resulted in several outstanding DCNN-based face recognition models, including DeepFace [33], DeepID [34][35][36][37], and FaceNet [38]. These models leverage CNNs to extract and classify facial features, achieving high accuracy rates in face recognition while also delivering good performance in terms of detection speed. Compared to traditional CNNs, DCNNs typically possess more convolutional layers and a larger parameter space.

3.1 DeepFace

DeepFace is a seminal model in the field of deep learning for face recognition, achieving a recognition rate of 97.53% on the Labelled Faces in the Wild (LFW) dataset, which is comparable to human performance [37]. DeepFace's face recognition process can be divided into four steps: face detection, face alignment, feature extraction, and classification. Firstly, the model identifies the human face and six fiducial points, followed by texture mapping using local binary pattern histograms to extract the corresponding features and derive a 3D face model. Radiative changes are then made based on the six fiducial points to obtain the corresponding 67 facial landmark points. Finally, the corresponding 3D face is obtained through triangulation. While this alignment approach is more complicated than subsequent deep learning-based approaches, it is beneficial for extracting more effective features. After alignment, DeepFace employs a DCNN to extract facial features. It utilizes two shared convolutional layers, three unshared convolutional layers, and two fully connected layers to process the image. Unshared convolution kernels are used to reduce information loss since different facial regions have varying local statistical features [39]. Finally, SoftMax is used for classification, and Dropout is incorporated to alleviate model overfitting. However, the 3D alignment approach used in DeepFace is computationally complex, and the model recognition speed drops to around 5 images per second due to the three unshared convolutional layers, while the parameter quantity is doubled. Furthermore, the classifier architecture of DeepFace requires different training for different data inputs to maintain accuracy, which reduces the model's usability.

3.2 DeepID

The DeepID model for face recognition includes detection, alignment, feature extraction, and classification. It uses multi-channel, multi-scale, and multi-region segmentation to detect faces and fiducial points before inputting data into the model [40]. The DeepID architecture has four convolutional layers with a pooling layer designed after the first three layers. Celeb-Faces is used to train the DeepID model, and the joint Bayesian model is used to improve accuracy [41]. DeepID2 abandons the SoftMax classifier and uses multiple Bayesian classifiers fused into a single ensemble via Support Vector Machine (SVM) [42]. DeepID2+ builds upon DeepID2 by connecting the DeepID layer with the max-pooling at each layer, increasing the dimensionality of the last layer, and fusing two datasets during training. DeepID2+ is more robust to occluded faces and achieves an accuracy of about 99% on the LFW dataset even with 10% random occlusion [43]. The deeper NNs of VGG and GoogLeNet models did not produce better results than DeepID3.

3.3 FaceNet

In 2015, Google developed the FaceNet model for face recognition. It maps human faces to Euclidean space using a CNN to extract features from the input image and calculate the Euclidean distances of the features, surpassing DeepID and achieving high accuracies on LFW and YouTube databases. The quality of the model is directly affected by the triplet loss function, making triplet selection crucial for improving performance. A lightweight FaceNet model is proposed in [44] with nearly identical accuracies to the original model but lower computational burden. In [45], the center and SoftMax losses are monitored jointly, making the learned features more discriminative and generalizable with faster convergence and requiring fewer sample sizes. As stated in [46], Vu et al, combined FaceNet and SVM for face recognition, achieving a high recognition rate on the LFW dataset.

3.4 Baidu

Baidu proposed a method for human face recognition in 2015 that combined Deep CNNs on multi-patch with Deep Metric Learning, achieving a 99.77% accuracy on the LFW dataset [47]. This approach involves segmenting an aligned face image into multiple overlapping patches and inputting them separately into the same network for training. The multi-patch technique's extracted features are more robust under complex conditions such as posture, occlusion, and expression. The features' dimensionality is then reduced to 128 using metric learning supervised by triplet loss. The DCNN structure on multi-patch includes nine convolution layers and a SoftMax layer at the end for supervised multi-class learning. The network's input is a 2D aligned RGB face image, with Pooling and Normalization layers between some convolution layers. The same structure is applied on overlapped image patches centered at different landmarks on the face region.

3.5 SphereFace

SphereFace is a face recognition algorithm that was proposed to overcome the challenges of complex scenarios [48][49]. It is the first algorithm to transform the feature space into a hypersphere angular feature space. The A-SoftMax and AM-SoftMax methods were developed to improve SphereFace's limitations. The AM-SoftMax method reduces intra-class spacing by subtracting a value from the cosine value but does not expand inter-class spacing. Improvements were made to AM-SoftMax by subtracting $m/2$ from the normalized objective function $\cos\theta_{y_i}$ and adding $m/2$ to the non-objective function $\cos\theta_j$ [50]. The central loss function, which combines angular margin loss and central loss, outperforms the traditional SoftMax under the same network architecture [51].

3.6 CosFace

In 2018, researchers proposed the CosFace algorithm, which is based on the Large Margin Cosine Loss (LMCL) approach [52]. This algorithm aims to address the issues with SphereFace and introduces LMCL, which is formulated as shown in equation (3) below.

$$-\frac{1}{N}\sum_{i} log \frac{e^{s\left(\cos(\theta_{y_i})-m\right)}}{e^{s\left(\cos(\theta_{y_i})-m\right)}+\sum_{j\neq y_i} e^{s\cos(\theta_j)}} \tag{1}$$

The decision boundary of A-SoftMax:

$$C_1{:}\cos(m\theta_1) \geq \cos(\theta_2)$$
$$C_2{:}\cos(m\theta_2) \geq \cos(\theta_1)$$

was changed into:

$$C_1{:}\cos(\theta_1) \geq \cos(\theta_2) + m$$
$$C_2{:}\cos(\theta_2) \geq \cos(\theta_1) + m$$

The LMCL can restrict the cosine value by adding margin and expand the inter-class angular distance by modifying the m value. The accuracies of CosFace on LFW and YouTube Face (YTF) datasets are 99.73% and 97.6%, respectively, both surpassing those of SphereFace.

3.7 ArcFace

To improve the recognition ability of face recognition models and stabilize the training process, a novel additive angular margin loss was proposed by [53] in 2019. By training with this loss function, they obtained the ArcFace model. Compared to other loss functions, ArcFace directly maximizes the classification boundary in the angular space and moves the penalty term m from outside to inside of the cosine function. This approach poses stricter classification requirements and further enhances the classification ability of face recognition networks [54]. Different loss functions were compared in [55] for binary classification, and ArcFace was found to have a constant linear angular margin and better discriminative power in face recognition than SphereFace and CosFace. It is also computationally lightweight and easy to implement. ArcFace was further improved in [56] with a dynamic adaptive scaling factor and a style attention mechanism. The fair loss was proposed in [57] to address the class imbalance problem in face recognition, where some classes have more samples than others, by allowing each class to learn an appropriate adaptive margin. The fair loss outperforms other methods on all three datasets.

3.8 Summary

The development of DCNNs has greatly advanced the field of face recognition. DeepFace, the first model to apply DCNN to face recognition, uses a procedure of detection–alignment–extraction–classification and achieves impressive results on the LFW dataset, providing a reference for subsequent face recognition research. However, the computational complexity of 3D alignment and the large number of network parameters limit the model's usability. In 2014, the DeepID model was proposed with a multi-scale, multi-channel approach and extended the dataset, improving the accuracy on LFW to 97.45%. DeepID, however, uses SoftMax, which is ineffective in representing features. Later models such as DeepID2 and DeepID2+ use joint Bayesian classification and increase feature dimensionality to improve training results and robustness to occlusion. FaceNet uses neither complex alignment nor SoftMax classification, instead normalizing extracted features and calculating Euclidean distance using Triplet Loss. However, an inappropriate selection strategy for triples can lead to overfitting. Baidu integrates multi-patch and metric learning for better robustness in complex scenarios. A novel loss function based on center loss was proposed in 2016, which when combined with SoftMax loss, improves network classification ability and face recognition performance. SphereFace, CosFace, and ArcFace introduce the multiplicative angular margin, the additive cosine angular margin loss, and the additive angular margin to the loss function, respectively, narrowing the intra-class spacing and expanding the inter-class spacing to enhance SoftMax's classification ability. While current loss functions are impressive, there remains ample room for improvement. In [58], a combination of guided image filters and CNNs is proposed to reduce the effects of lighting, pose, and expression on faces. In [59], a face recognition approach based on non-subsampling shearlet transform (NSST), CNN, and SVM is proposed. In [60], a hybrid ConvNet approach is used for face validation to learn face similarity between image pairs. Table 2 lists the accuracies of face recognition models on the LFW, YTF (YouTube Faces), CFP-FP (Celebrities in Frontal-Profile in the Wild), and AgeDB-30 datasets, where "Data" indicates the number of parameters in the model.

Table 2. Comparison of accuracies of different face recognition models

Method	Years	Data (M)	LFW (%)	YTF (%)	CFP-FP (%)	AgeDB-30 (%)
DeepFace	2014	4	97.35	91.4	–	–
DeepID	2014	0.2	97.45	–	–	–
DeepID2	2014	0.3	99.15	–	–	–
DeepID2+	2015	0.3	99.47	93.2	–	–
DeepID3	2015	0.3	99.53	–	–	–
FaceNet	2015	200	99.63	95.1	–	–
Baidu	2015	1.3	99.13	–	–	–
[61]	2016	0.7	99.28	94.9	–	–
SphereFace	2017	0.5	99.42	95.0	94.17	97.30
[45]	2018	0.7	99.31	–	–	–
CosFace	2018	5	99.73	97.6	94.4	97.91
ArcFace	2019	5.8	99.83	98.02	95.56	95.15
[57]	2020	5.2	99.15	–	98.85	91.24

As shown in Table 2, ArcFace showed the best performance with an accuracy of 99.83% on the LFW dataset and 98.02% on the YTF dataset, respectively. The method proposed in [57] obtained the highest accuracy of 98.85% on the CFP-FP dataset. Meanwhile, CosFace had the highest accuracy of 97.91% on the AgeDB-30 dataset.

4 DISCUSSIONS AND RECOMMENDATIONS

In term of image classification, the LeNet-5 is a classic CNN architecture that is not as powerful as some of the newer architectures, but it is still a good choice for simple image recognition tasks. It is recommended for beginners who want to understand the basics of CNNs. The AlexNet and VGG are good choice for general image recognition tasks, especially to deal with a large dataset. It is easy to understand and implement, making it a popular choice for academic research. However, it can be computationally expensive. A CPU is required, as otherwise the experimental results may vary and not be significant. The GoogLeNet is recommended for tasks where computational efficiency is crucial. The ResNet-152 is a deeper version of ResNet and is suitable for complex image recognition tasks, especially when dealing with large datasets. But, it can be computationally expensive as well. The MobileNet is designed for mobile and embedded devices, offering a good trade-off between accuracy and model size. The MobileNetv2 offers better performance and efficiency, providing higher accuracy at the cost of a slightly larger model size. It is recommended for mobile applications where a balance between accuracy and size is needed and the MobileNetv3 is recommended for applications that require real-time performance on mobile devices. The ShuffleNet offers higher accuracy at the expense of a larger model size. The ShuffleNetv2 is recommended for resource-constrained scenarios. The EfficientNet is the baseline version and is recommended as a starting point for many image recognition tasks. The EfficientNet-B6 is recommended when you need high accuracy with constrained resources.

In terms of face recognition algorithms, the DeepFace and DeepID models are both excellent choices for face verification tasks, especially in controlled environments and simple face recognition projects. However, DeepID outperforms DeepFace in terms of power and accuracy for face identification tasks. Additionally, DeepID3, an advanced version of the DeepID series, offers improved performance and robustness, making it well-suited for tasks that demand high accuracy and can handle variations in facial appearance. For highly accurate face recognition tasks, FaceNet is widely recognized for its ability to learn highly discriminative face embeddings, making it a recommended choice. Despite Baidu's face recognition technology being known for its efficiency and accuracy, specific details about its algorithms are not publicly available. Therefore, Baidu may not be suitable for research purposes. For tasks requiring a high level of robustness and accuracy, SphereFace is a suitable option. CosFace is a recommended choice for tasks that prioritize both accuracy and robustness. Lastly, ArcFace is the ideal algorithm for tasks where achieving high accuracy and robustness is critical.

5 CONCLUSION AND FUTURE WORKS

CNNs are highly regarded for their local connection and weight sharing capabilities, making them vital in computer vision and image recognition. Their importance will continue to be a topic of research for the foreseeable future, and face recognition technology based on DNN will become increasingly sophisticated. Researchers have been focusing on modifying loss functions to improve the generalization ability of networks and increase model performance. Despite these efforts, several challenges still persist. In complex scenarios, human faces are susceptible to factors such as posture, expression, illumination, and occlusion, which may greatly reduce the recognition rate. In the future, the primary challenge for face recognition will be capturing subtle inter-class variations amidst significant interference from intra-class changes.

With the continuous development of science and technology, face recognition research is likely to head in the following directions in the future:

- Lightweight face recognition: Some bulky networks require a significant amount of memory and computational power, making it impractical to apply them on mobile devices. Although researchers have been working on lightweight face recognition, improving its efficiency and accuracy is still highly necessary.
- Noise label learning: In the process of collecting large-scale face data, there is often a label noise problem. Researchers are exploring ways to build clean datasets by denoising or learning noise-robust face representations. However, these approaches are often impacted by network model ability and cannot be flexibly applied in actual scenarios. Noise label learning remains an unsolved issue in face recognition technology.
- Unlabeled face learning: With the increase in data size, manual labeling has become too expensive. There are many unlabeled face datasets, and another future research direction is to explore ways to perform face recognition using unlabeled face datasets while retaining high accuracy.

This study summarizes the advances, insights, and future prospects of DCNN-based face recognition technology based on representative face recognition models over the years. The background and general face recognition process are initially described, followed by an outline of classic CNN models. Next, the DCNN-based face recognition models are reviewed, and suggestions for improvements to the models are provided. Finally, the difficulties and challenges encountered by face recognition are analyzed, and future development directions are proposed.

6 ACKNOWLEDGMENT

This work was supported in part by the University Malaysia Sabah under SDN Grant SDN51-2019, as well as the Handan Municipal Bureau of Science and Technology, Hebei Province, China under Grant 21422031252. The authors also wish to thank Universiti Malaysia Sarawak for the financial support of this project.

7 REFERENCES

[1] W. Liang, "Research on face recognition algorithm based on independent component analysis," Master's thesis, Xi'an University of Science and Technology, Xi'an, China, 2012.

[2] S. K. Chung, K. O. Chin, M. H. A. Hijazi, and M. M. Singh, "Smart-Hadir – Mobile based attendance management system," *International Journal of Interactive Mobile Technologies*, vol. 15, no. 14, pp. 4–16, 2021. https://doi.org/10.3991/ijim.v15i14.22677

[3] H. Du, H. Shi, D. Zeng, X. P. Zhang, and T. Mei, "The elements of end-to-end deep face recognition: A survey of recent advances," *ACM Computing Surveys (CSUR)*, 2020. https://doi.org/10.1145/3507902

[4] M. S. M. Suhaimin, M. H. A. Hijazi, C. S. Kheau, and C. K. On, "Real-time mask detection and face recognition using eigenfaces and local binary pattern histogram for attendance system," *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 2, pp. 1105–1113, 2021. https://doi.org/10.11591/eei.v10i2.2859

[5] S. Evan, L. Jonathan, and D. Trevor, "Fully convolutional networks for semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 640–651, 2017. https://doi.org/10.1109/CVPR.2015.7298965

[6] N. Matnoor and N. Suaib, "Review on facial expression modeling," *Bulletin of Electrical Engineering and Informatics*, vol. 11, no. 2, pp. 779–784, 2022. https://doi.org/10.11591/eei.v11i2.3558

[7] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998. https://doi.org/10.1109/5.726791

[8] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," *Advanced in Neural Information Processing Systems*, vol. 25, 2012. https://doi.org/10.1145/3065386

[9] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv e-prints, arXiv:1409.1556, 2014.

[10] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, … and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9. https://doi.org/10.1109/CVPR.2015.7298594

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778. https://doi.org/10.1109/CVPR.2016.90

[12] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, … and A. Hartwig, “MobileNets: Efficient convolutional neural networks for mobile vision applications,” arXiv e-prints, arXiv:1704.04861, 2017.

[13] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “MobileNetV2: Inverted residuals and linear bottlenecks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510–4520. https://doi.org/10.1109/CVPR.2018.00474

[14] A. Howard, M. Sandler, G. Chu, L. C. Chen, B. Chen, M. Tan, … and Q. V. Le, “Searching for MobileNetV3,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1314–1324. https://doi.org/10.1109/ICCV.2019.00140

[15] X. Zhang, X. Zhou, M. Lin, and J. Sun, “ShuffleNet: An extremely efficient convolutional neural network for mobile devices,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6848–6856. https://doi.org/10.1109/CVPR.2018.00716

[16] N. Ma, X. Zhang, H. T. Zheng, and J. Sun, “ShuffleNet V2: Practical guidelines for efficient CNN architecture design,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 116–131. https://doi.org/10.1007/978-3-030-01264-9_8

[17] M. Tan and Q. V. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” arXiv preprint arXiv:1905.11946, 2019.

[18] M. Tan and Q. V. Le, “EfficientNetV2: Smaller models and faster training,” arXiv preprint arXiv:2104.00298, 2021.

[19] Z. Fouad, M. Alfonse, M. Roushdy, and A. M. Salem, “Hyper-parameter optimization of convolutional neural network based on particle swarm optimization algorithm,” *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 6, pp. 3377–3384, 2021. https://doi.org/10.11591/eei.v10i6.3257

[20] S. Shindo, T. Goto, T. Kirishima, and K. Tsuchida, “An optimization of facial feature point detection program by using several types of convolutional neural network,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 16, no. 2, pp. 827–834, 2019. https://doi.org/10.11591/ijeecs.v16.i2.pp827-834

[21] B. J. Lin, Z. M. Geng, Y. Hong, and X. F. Li, “Convolutional neural networks for textile and apparel image applications,” *Journal of Beijing Institute of Fashion (Natural Science Edition)*, vol. 41, no. 1, pp. 92–99, 2021. https://doi.org/10.16454/j.cnki.issn.1001-0564.2021.01.014

[22] Z. Yu, Y. Dong, J. Cheng, M. Sun, and F. Su, “Research on face recognition classification based on improved GoogleNet,” *Security and Communication Networks*, vol. 2022, pp. 1–6, 2022. https://doi.org/10.1155/2022/7192306

[23] C. Jing, Z. Haopeng, C. Kim On, E. G. Moung, and P. Anthony, “Face recognition based on deep convolutional support vector machine with bottleneck attention,” *IAENG International Journal of Computer Science*, vol. 49, no. 4, pp. 1284–1296, 2022.

[24] Y. Li, Z. Hao, and H. Lei, “A review of research on convolutional neural networks,” *Computer Applications*, vol. 36, no. 9, pp. 2508–2515+2565, 2016. https://doi.org/10.11772/j.issn.1001-9081.2016.09.2508

[25] Z. Z. Yang, N. Kuang, L. Fan, and B. Kang, “A review of image classification algorithms based on convolutional neural networks,” *Signal Processing*, vol. 34, no. 12, pp. 1474–1489, 2018. https://doi.org/10.16798/j.issn.1003-0530.2018.12.009

[26] C. Chen and F. Qi, “A review of the development of convolutional neural networks and their applications in computer vision,” *Computer Science*, vol. 46, no. 3, pp. 63–73, 2019. https://doi.org/10.11896/j.issn.1002-137X.2019.03.008

[27] P. F. Ke, M. G. Cai, and T. Wu, “Face recognition algorithms based on improved convolutional neural networks with integrated learning,” *Computer Engineering*, vol. 46, no. 2, pp. 262–273, 2020. https://doi.org/10.19678/j.issn.1000-3428.0053576

[28] L. H. Huang, Z. C. Kang, C. F. Zhang, and T. Cheng, “A lightweight convolutional neural network-based face recognition method,” *Journal of Hunan University of Technology*, vol. 33, no. 2, pp. 43–47, 2019. https://doi.org/10.3969/j.issn.1673-9833.2019.02.008

[29] W. X. Wang, X. Zhou, X. H. He, L. B. Qi, and Z. Y. Wang, "Face expression recognition based on improved MobileNet networks," *Computer Applications and Software*, vol. 37, no. 4, pp. 137–144, 2020. https://doi.org/10.3969/j.issn.1000-386x.2020.04.023

[30] Q. Cai, C. Peng, and X. Shi, "A lightweight face recognition algorithm based on MobieNetV2," *Computer Applications*, vol. 40, no. 1, pp. 65–68, 2020. https://doi.org/10.11772/j.issn.1001-9081.2019122219

[31] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 8, pp. 2011–2023, 2020. https://doi.org/10.1109/CVPR.2018.00745

[32] F. Y. Song, L. M. Wu, G. Z. Zheng, and X. Y. He, "Structural pruning optimization based on MobileNetV3," *Automation and Information Engineering*, vol. 40, no. 6, pp. 20–25, 2019.

[33] Y. Taigman, M. Yang, M. A. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1701–1708. https://doi.org/10.1109/CVPR.2014.220

[34] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," *Advances in Neural Information Processing Systems*, vol. 27, 2014.

[35] Y. Sun, D. Liang, X. Wang, and X. Tang, "DeepID3: Face recognition with very deep neural networks," ArXiv, abs/1502.00873, 2015.

[36] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1891–1898. https://doi.org/10.1109/CVPR.2014.244

[37] Y. Sun, X. Wang, and X. Tang, "Deeply learned face representations are sparse, selective, and robust," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2892–2900. https://doi.org/10.1109/CVPR.2015.7298907

[38] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 2015, pp. 815–823. https://doi.org/10.1109/CVPR.2015.7298682

[39] J. Y. Zhou and Y. M. Zhao, "A review of convolutional neural networks for image classification and target detection applications," *Computer Engineering and Applications*, vol. 53, no. 13, pp. 34–41, 2017. https://doi.org/10.3778/j.issn.1002-8331.1703-0362

[40] W. D. Zhang, Y. L. Xu, J. C. Ni, S. P. Ma, and H. Shi, "Image target recognition algorithm based on multiscale chunked convolutional neural network," *Computer Applications*, vol. 36, no. 4, pp. 1033–1038, 2016. https://doi.org/10.11772/j.issn.1001-9081.2016.04.1033

[41] Y. Liu, "Research and implementation of large-scale face recognition technology based on improved DeepID," M.S. thesis, Dept. Comp., Jilin University, Changchun, China, 2017.

[42] J. Jiang, "Research on face recognition technology based on deep learning," M.S. thesis, Dept. Comp., Harbin Institute of Technology, Harbin, China, 2019.

[43] G. Li, "A review of patent technology for face recognition," *Inventions and Innovations (Vocational Education)*, vol. 3, pp. 73–75, 2019.

[44] X. Xu, M. Du, H. Guo, J. Chang, and X. Zhao, "Lightweight FaceNet based on MobileNet," *International Journal of Intelligence Science*, vol. 11, no. 1, p. 16, 2021. https://doi.org/10.4236/ijis.2021.111001

[45] C. B. Yu, T. Tian, D. E. Xiong, and L. Y. Xu, "Face recognition under joint supervision of central loss and Softmax loss," *Journal of Chongqing University*, vol. 41, no. 5, pp. 92–100, 2018. https://doi.org/10.11835/j.issn.1000-582X.2018.05.012

[46] L. Vu, P. Trieu, and H. Nguyen, "Implementation of FaceNet and support vector machine in a real-time web-based timekeeping application," *IAES International Journal of Artificial Intelligence*, vol. 11, no. 1, pp. 388–396, 2022. https://doi.org/10.11591/ijai.v11.i1

[47] J. Liu, Y. Deng, T. Bai, Z. Wei, and C. Huang, "Targeting ultimate accuracy: Face recognition via deep embedding," arXiv preprint arXiv:1506.07310, 2015.

[48] Y. A. Zhang, H. Y. Wang, and F. Xu, "Face recognition based on deep convolutional neural networks with central loss," *Science Technology and Engineering*, vol. 17, no. 35, pp. 92–97, 2017. https://doi.org/10.3969/j.issn.1671-1815.2017.35.015

[49] R. Gong, D. Sheng, C. H. Zhang, and H. Su, "A lightweight and multi-pose face recognition method based on deep learning," *Computer Applications*, vol. 40, no. 3, pp. 704–709, 2020. https://doi.org/10.11772/j.issn.1001-9081.2019071272

[50] S. Zhang, Y. H. Gong, and J. Wang, "Development of deep convolutional neural networks and their applications in computer vision," *Journal of Computer Science*, vol. 42, no. 3, pp. 453–482, 2019. https://doi.org/10.11897/SP.J.1016.2019.00453

[51] Z. D. Li, Y. Zhong, M. Chen, and L. S. Wang, "Deep face recognition with combined angular residual loss and center loss," *Computer Applications*, vol. 39, no. 2, pp. 55–58, 2019.

[52] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "Cosface: Large margin cosine loss for deep face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5265–5274. https://doi.org/10.1109/CVPR.2018.00552

[53] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4690–4699. https://doi.org/10.1109/CVPR.2019.00482

[54] Q. Guo, Z. H. Wang, D. Fan, and H. J. Wu, "Design and implementation of a multiple face recognition system based on deep learning," *Journal of Inner Mongolia Agricultural University (Natural Science Edition)*, pp. 1–7, 2022. https://doi.org/10.16853/j.cnki.1009-3575.2022.02.015

[55] J. Wang, R. Liu, and Y. Hou, "Research on DeepID-based face detection and recognition algorithm," *Computer Knowledge and Technology*, vol. 14, no. 17, pp. 220–221, 2018. https://doi.org/10.14004/j.cnki.ckt.2018.1875

[56] Z. H. Zhang and R. Wang, "An improved face recognition method based on MobileFaceNet network," *Journal of Beijing University of Aeronautics and Astronautics*, vol. 46, no. 9, pp. 1756–1762, 2020. https://doi.org/10.13700/j.bh.1001-5965.2020.0049

[57] B. Liu, W. Deng, Y. Zhong, M. Wang, J. Hu, X. Tao, and Y. Huang, "Fair loss: Margin-aware reinforcement learning for deep face recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 10052–10061. https://doi.org/10.1109/ICCV.2019.01015

[58] Y. Singamsetti and N. Purnachand, "Convolutional neural network-based face recognition using non-subsampled shearlet transform and histogram of local feature descriptors," *IAES International Journal of Artificial Intelligence*, vol. 10, no. 4, pp. 1079–1090, 2021. https://doi.org/10.11591/ijai.v10.i4

[59] Y. Singamsetti and N. Purnachand, "An effective face recognition method using guided image filter and convolutional neural network," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 23, no. 3, pp. 1699–1707, 2021. https://doi.org/10.11591/ijeecs.v23.i3

[60] F. Zaman, J. Johari, and A. Yassin, "Learning face similarities for face verification using hybrid convolutional neural networks," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 16, no. 3, pp. 1333–1342, 2019. https://doi.org/10.11591/ijeecs.v16.i3

[61] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *European Conference on Computer Vision*, Springer, Cham, 2016, pp. 499–515. https://doi.org/10.1007/978-3-319-46478-7_31

8 AUTHORS

Jing Chi is an Assistant Professor in the School Information and Electrical Engineering, Hebei University of Engineering. She is now pursuing a PhD in computer science at University Malaysia Sabah (UMS). Her research interest generally falls under Computer Vision & Pattern Recognition, such as image processing, image segmentation, image classification, object detection, and vision-based learning (E-mail: chijing@hebeu.edu.cn).

Chin Kim On is currently working as an Associate Professor at the Universiti Malaysia Sabah in the Faculty of Computing and Informatics. His research interests are gaming AI, evolutionary computing, evolutionary robotics, artificial neural networks, image processing, agent technologies, evolutionary data mining, and IoT. He has led several projects related to artificial neuro-cognition for solving real world problems such as mobile based number plate detection and recognition, off-line handwriting recognition, item drop mechanism and auto map generation in gaming AI, as named a few (E-mail: kimonchin@ums.edu.my).

Haopeng Zhang is a master student in the School of Information and Electrical Engineering, Hebei University of Engineering, graduated from Hebei University of Economics and Business with a bachelor degree in Computer Science and Technology in June 2020, and is a member of CCF. His research interests are in the field of computer vision and pattern recognition, such as image processing, image classification, target detection, etc. (E-mail: zhanghaopengyyds@163.com).

Soo See Chai is presently working as a senior lecturer at Department of Computing and Software Engineering at Faculty of Computer Science and Information Technology (FCSIT) in Universiti Malaysia Sarawak (UNIMAS). Her research areas are GIS, Re-mote Sensing, AI and Image Processing (E-mail: sschai@unimas.my).