Proceedings of CONF-MLA 2024 Workshop: Securing the Future: Empowering Cyber Defense with Machine Learning and Deep Learning
DOI: 10.54254/2755-2721/110/2024MELB0107

Enhancing Facial Recognition: A Comprehensive Review of Deep Learning Approaches and Future Perspectives

Songze Zhu

School of Software, Harbin Institute of Information Technology, Harbin, China

fengsheng@ldy.edu.rs

Abstract. This paper presents a comprehensive review of the application of deep learning in facial recognition, covering fundamental aspects from neural network architectures to advanced training methods. It starts with an introduction to the basic architectures such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), including variations like Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks. The review extends to sophisticated architectures like AlexNet, GoogleNet, VGGNet, and ResNet, which have significantly pushed the boundaries of accuracy in face recognition tasks. The paper details the process of data preprocessing in face recognition, which involves critical steps such as face detection, alignment, and normalization to ensure uniformity in the input data, enhancing the accuracy of feature extraction. Various feature extraction methods are discussed, including CNN-based and Generative Adversarial Network (GAN)-based techniques, which have shown considerable promise in deriving complex facial features from raw images. Loss functions such as Euclidean distance loss, angular/cosine margin loss, and softmax loss variants are explored to understand their impact on enhancing the discriminative power of the facial recognition systems. The paper highlights the evolution from traditional models using eigenfaces and feature descriptors like Local Binary Patterns (LBP) to cutting-edge deep learning models that utilize deep identity features (DeepID) and triplet loss functions to improve recognition accuracy.

Keywords: Facial recognition, deep learning, neural network.

1. Introduction

Facial recognition technology has significantly evolved from the traditional use of photographic identification by law enforcement and businesses to a widespread application across various digital platforms. With the advent of ubiquitous cameras and mobile devices, the acquisition and dissemination of facial images have become part of everyday life, leading to increased interest and research in improving facial recognition technologies [1]. While facial recognition offers distinct advantages over other biometric systems, such as iris scans or fingerprints, due to its non-intrusive data collection methods, it also faces unique challenges. These challenges include variations in appearance due to makeup, lighting, and occlusions, which can significantly affect performance and reliability [2].

Deep learning has revolutionized the field of facial recognition, providing substantial improvements in accuracy and robustness. Techniques like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have become foundational in developing systems that can effectively handle the complex variability of facial features encountered in real-world scenarios [3, 4]. Innovations in

DOI: 10.54254/2755-2721/110/2024MELB0107

network architectures, including AlexNet, GoogleNet, VGGNet, and ResNet, have set new standards for what can be achieved in terms of deep learning capabilities, pushing the boundaries of face recognition technology further [5]. These advancements have been supported by comprehensive studies on different neural network configurations and training methods that optimize performance across varied conditions and datasets.

This paper provides a comprehensive review of how deep learning techniques are applied to enhance facial recognition systems. The discussion begins with an overview of fundamental deep learning concepts, focusing on neural network architectures and their specific applications in facial recognition. It explores the various stages of facial recognition processing, including data preprocessing, feature extraction, and the application of sophisticated loss functions to refine accuracy. The paper evaluates both traditional and contemporary models, highlighting the progression from elementary techniques using eigenfaces and feature descriptors to advanced deep learning models that utilize deep identity features and complex network architectures. By examining these elements in detail, the paper aims to present a clear picture of the current landscape of facial recognition technologies and provide insight into potential future research directions that could further improve the efficacy and applicability of these systems.

2. Fundamentals of deep learning

2.1. Overview

A sub-field of machine learning known as "deep learning" makes use of algorithms with intricate architecture or several processing layers composed of various non-linear transformations, which aims to thoroughly abstract input. Face recognition algorithms that rely on deep learning, acquire the capacity to extract features in an end-to-end fashion and apply those features to categorization.

2.2. Neural network architecture

2.2.1. Basic architecture. Convolutional Neural Network (CNN). Layers for input, convolution, subsampling, complete connection, and output make up the fundamental structure. Performing numerous convolution and pooling procedures on the input picture is the fundamental notion of CNN, which is to extract image information [3]. Stacking layers allows one to ultimately acquire the mapping connection between the classifier's input and output.

Pooling layer: As a nonlinear dimensionality reduction form inside the CNN, pooling layer reduces the size of the feature map while retaining the basic features. Among the various pooling methods, maximum pooling is particularly effective for subsampling operations [4].

Fully connected layer (FC): FC layers perform the final data analysis in a neural network [5]. As show in the figure 1.

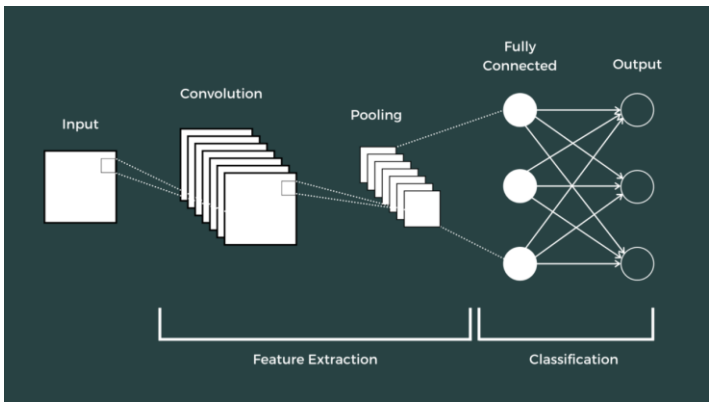

Figure 1. Architecture of CNN (Photo credit: Original).

DOI: 10.54254/2755-2721/110/2024MELB0107

Recurrent Neural Network (RNN). Unlike standard neural networks that handle data points separately, RNNs account for the order in which data points occur, which is beneficial for tasks with time-based data [6]. Figure 2 depicts a simple recurrent neural network, where the internal memory h_t is computed using the following quation:

$$h_t = g(Wx_t + Uh_t + b) \tag{1}$$

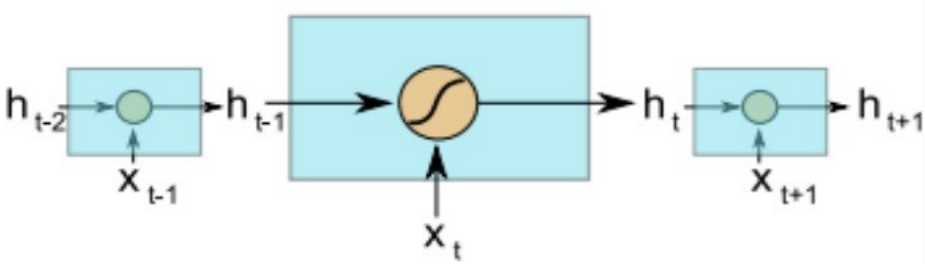

Figure 2. Architecture of RNN (Photo credit: Original).

As show in the figure 2. Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU). During extended sequence training, gradient disappearance and explosion are two common problems that long short-term memory (LSTM) is intended to address [7, 8].

Three gates (an input gate, an output gate, and a forget gate) are introduced by LSTM in contrast to the conventional RNN (recurrent neural network). The output gate, finally, regulates how the cell state affects the output [9].

As show in the figure 3 and 4. The update equations for the LSTM unit are expressed by Equation:

$$\begin{aligned} h^{(t)} &= g_0^{(t)} f_h(s^{(t)}) \\ s^{(t-1)} &= g_f^{(t)} s^{(t-1)} + f_s(wh^{(t-1)}) + uX^{(t)} + b \\ g_i^{(t)} &= sigmoid(w_i h^{(t-1)} + u_i X^{(t)} + b_i) \\ g_f^{(t)} &= sigmoid(w_f h^{(t-1)} + u_f X^{(t)} + b_f) \\ g_o^{(t)} &= sigmoid(w_o h^{(t-1)} + u_o X^{(t)} + b_o) \end{aligned} \tag{2}$$

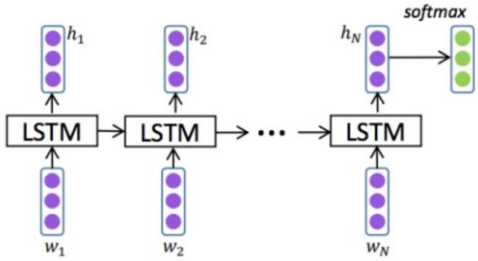

Figure 3. The architecture of LSTM model (Photo credit: Original).

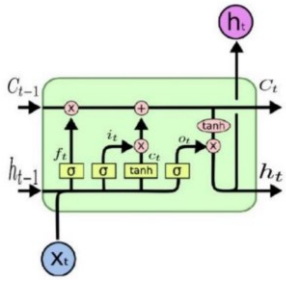

Figure 4. The architecture of gates in LSTM (Photo credit: Original).

DOI: 10.54254/2755-2721/110/2024MELB0107

2.2.2. Advanced architecture. AlexNet. AlexNet, a deep Convolutional Neural Network (CNN), secured victory in the 2012 edition of ILSVRC. It also employs a 1000-class softmax output, dropout for regularization, ReLU activation functions, and data augmentation to enhance its performance [10]. As show in the figure 5.

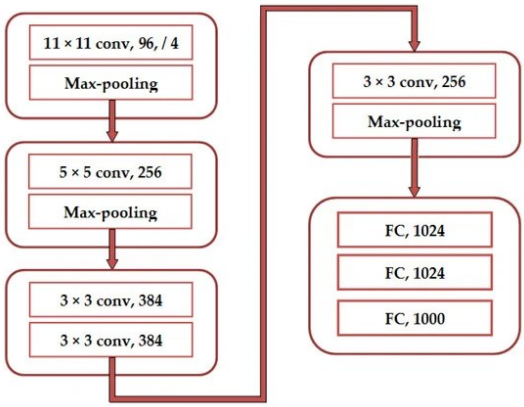

Figure 5. AlexNet architecture. FC: fully connected layers. conv: convolution (Photo credit: Original).

GoogleNet. This architecture was designed to reduce computational demands compared to traditional CNNs. It introduced the "inception module," which utilizes various kernel sizes to create receptive fields of different sizes [11]. Within these modules, multiple convolutions of sizes 1x1, 3x3, and 5x5, along with 3x3 max-pooling, operate in parallel on the input [12]. As show in the figure 6.

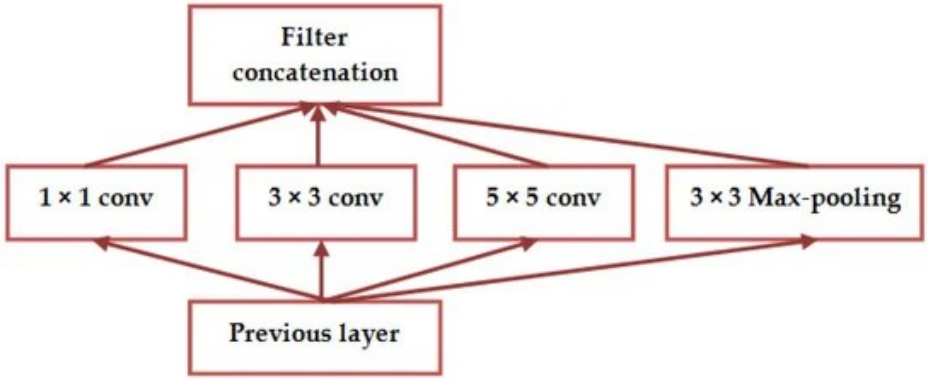

Figure 6. GoogleNet architecture (Photo credit: Original).

ResNet. In 2015, He et al. introduced ResNet, a groundbreaking deep learning architecture, which won the ILSVRC competition that year. This model was designed to handle extremely deep neural networks by incorporating "shortcut connections" and employing batch normalization [13, 14]. It enabled the training of networks with a range of depths, including 34, 50, 101, 152, and even 1202 layers. As show in the figure 7.

DOI: 10.54254/2755-2721/110/2024MELB0107

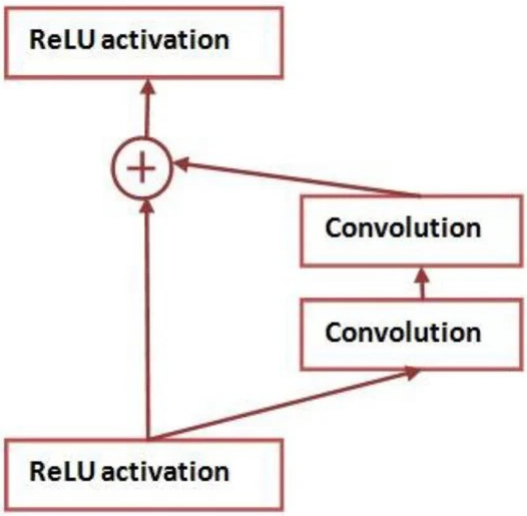

Figure 7. ResNet architecture (Photo credit: Original).

VGGNet. At the ILSVRC-2014 competition, Simonyan et al. studied the influence of network depth on image recognition accuracy [15]. By expanding VGGNet to 16 to 19 weight layers, they improved the VGGNet's ability to fit complex nonlinear mappings, demonstrating the benefits of deeper architectures. As show in the figure 8.

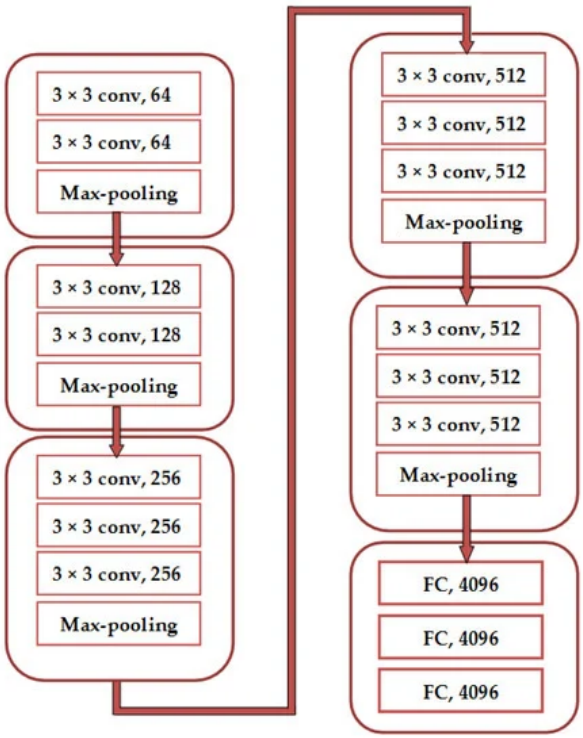

Figure 8. VGGNet architecture (Photo credit: Original).

2.3. Training methods

Typically, backpropagation is used in deep learning training to decrease the loss function through network weight adjustments. To increase training efficiency, other optimization techniques are employed, such as stochastic gradient descent (SGD), Adagrad, Adadelta, etc.

DOI: 10.54254/2755-2721/110/2024MELB0107

3. Deep learning methods in face recognition

3.1. Data preprocessing in face recognition

Data preprocessing in face recognition includes steps such as face detection, alignment, and normalization. Face alignment adjusts the detected faces to a uniform position and Angle for comparison and feature extraction [16]. Normalization eliminates scale, illumination and color differences between different images, making feature extraction and comparison fairer and more accurate. During this phase, techniques such as pre-processing can eliminate undesirable elements like noise, blur, inconsistent lighting, and shadows. With the refined, clear facial images obtained, we can then proceed to the feature extraction phase.

3.2. Feature extraction

3.2.1. CNN-based method. Directly use CNN to extract face features, and learn the abstract features of the face through multiple convolution and pooling operations.

In the actual scene, the face may have various sizes and possible positions, which is difficult to detect. Some studies have also adopted a multi-branch CNN architecture to extract features at different levels. Other optimization techniques, including stochastic gradient descent (SGD), Adagrad, Adadelta, etc., are used to improve training efficiency.

3.2.2. GAN-based method. A discriminant network and a generative network make up the generative adversarial network. The generation network's input consists of random samples drawn from latent space, and the output it produces must closely resemble the actual sample found in the training set. The actual sample serves as both the discriminant network's input and the produced network's output. The discriminant network should be fooled as much as feasible by the generating network. With continual parameter adjustments, the two networks operate in opposition to one another. By controlling the generation of different facial attributes (such as age, gender, expression), training samples with specific characteristics can be generated, thereby improving the generality of the model.

In addition, it can be used to reinforce the data to improve the robustness of the model in the face of noise and variations. However, the convergence theory of GAN is still under study. In fact, GAN often encounters the gradient disappearance/explosion problem.

3.3. Loss functions

3.3.1. Euclidean distance loss.

$$\mathrm{D}(\mathrm{x},\mathrm{y}) = (\mathrm{x}-\mathrm{y})^{\mathrm{T}}(\mathrm{x}-\mathrm{y}) = ||\mathrm{x}-\mathrm{y}||_2 \tag{3}$$

The contrastive loss a Euclidean distance-based loss function. It can be defined as:

$$L_{contrastive} = \sum_{i=1}^{p}((1-y^i)\frac{1}{2}||f(x_1^i)-f(x_2^i)|| + y^i\frac{1}{2}(max(0,m-||f(x_1^i)-f(x_2^i)||))^2 \tag{4}$$

When representing input pairs$(\mathrm{x}_1,\mathrm{x}_2)^{\mathrm{i}}$, yi = 0 and different inputs yi = 1 are used to represent their similarity. The boundary m is a positive value that places the boundary around f(x).

The triplet loss is a method that relies on Euclidean distance to measure loss.

$$L_{triplet} = \sum_{i=1}^{N}(||f(x_i^a)-f(x_i^p)|| - ||f(x_i^a)-f(x_i^n)|| + m) \tag{5}$$

The triplet loss function focuses on carefully chosen triplets of images $(f(x_i^a), f(x_i^p), f(x_i^n))$: The parameter m, representing the minimum separation.

DOI: 10.54254/2755-2721/110/2024MELB0107

3.3.2. Angle / cosine marginal loss. Enhancing the angular boundary's discriminative power can be achieved by directly employing the angular margin as a distance measure. A generic loss function based on angular distance can be formulated as follows:

$$L_{angular} = \frac{1}{N}\sum_{i=1}^{N} -log(\frac{e^{||x_i||cos(m\theta_{y_i,i})}}{e^{||x_i||cos(m\theta_{y_i,i})}+\sum_{j\neq y_i} e^{||x_i||cos(\theta_{y_i,i})}}) \quad (6)$$

Where $\theta_{y_i,i}$ is restricted in $[0,\frac{\pi}{\mathrm{m}}]$.

3.3.3. Softmax loss and its variants. The cross entropy (CE) technique was developed for an adaptive process aimed at calculating the likelihood of infrequent occurrences within intricate probabilistic systems.

$$\mathrm{CE}(\mathrm{g},\mathrm{p}) = -\sum_{\mathrm{j}} \mathrm{g_j}\log(\mathrm{p_j}) \quad (7)$$

Where g_j is the label of class j.

The L2-softmax loss incorporates an L2 normalization to the feature vectors, ensuring that they are confined to a hypersphere with a constant radius [17]:

$$L_{L2-softmax} = \frac{1}{N}\sum_{i=1}^{N} -log(\frac{e^{W_{y_i}^T f(x_i)+b_{y_i}}}{\sum_{j=1} e^{W_j^T f(x_i)+b_j}}) \quad (8)$$

$$\mathrm{s.t.}\,||\mathrm{f}(\mathrm{x_i})|| = \alpha \quad (9)$$

This approach ensures that both high-quality and low-quality facial images receive equal consideration, as all feature vectors are normalized to have the same L2-norm. Additionally, it enhances the verification accuracy by encouraging features from the same individual to cluster closely together and those from different individuals to be more distant in the normalized feature space.

3.4. Face recognition models

3.4.1. Traditional model. The eigenfaces method uses an efficient representation of a PCA face image, combined with a standard face image (feature map), a low-weight recombination of each image can approximate any face image [18]. By projecting an image of a face onto a map with a feature volume, you can calculate the weight of what kind of face represents.

Feature descriptors such as local binary pattern (LBP)] and oriented gradient histogram (HOG) are also used for face recognition [19].

Hann et al. introduced a revolutionary facial image representation using the local texture descriptor LBP. In this way, the facial image is segmented into several segments, combined with histogram, extracting the distribution of LBP features and expanding the descriptors used for facial recognition.

3.4.2. Deep learning model. The DeepID series represents an early application of deep learning in facial recognition, significantly enhancing accuracy through the extraction of facial depth features. The concept of deep identity features (DeepID) captures high-level facial representations, derived from the deepest layers of convolutional networks, capable of distinguishing around 10,000 different identities across all training data.A fundamental challenge in facial recognition is developing a feature representation that minimizes intra-individual differences while maximizing inter-individual variations. This issue is addressed through Deep Identification Verification features, known as DeepID2, which are trained using deep convolutional neural networks (CNNs) with dual supervisory signals. The identification signal sharpens distinctions between individuals by differentiating DeepID2 features across identities, while the verification signal diminishes variability within the same individual by aligning DeepID2 features from identical identities. The synergistic effect of these signals results in a

DOI: 10.54254/2755-2721/110/2024MELB0107

feature set more effective than those derived from individual signals alone. Following this, the DeepID2+ model was introduced, advancing its predecessor's capabilities. Inspired by architectures such as GoogLeNet and VGGNet, it incorporates stacked convolutions and inception layers, trained with both facial identification and verification signals applied at final and intermediate feature extraction stages. Additionally, the 2015 introduction of FaceNet by Schroff et al. marked a significant advancement with its employment of a triplet loss function, which facilitates the creation of distinctive facial feature representations using a deep convolutional network. This model, trained on a dataset containing millions of facial images, produces a 128-dimensional feature representation. SphereFace introduces another pivotal development with its Angular Margin Softmax Loss, which, unlike traditional softmax loss functions that assume a uniform distribution of classes, increases the angular margin between classes to enhance stability and accuracy in face recognition, even amidst input variations. CosFace's Large Margin Cosine Loss (LMCL) eliminates the impact of radial feature differences through L2 normalization of features, enhancing the discriminatory capacity of the model. ArcFace builds on this by optimizing the additive angular margin and feature vector normalization, focusing on maximizing classification boundaries directly in the angular space θ, as opposed to the cosine space in CosFace, thereby directly enhancing angular differentiation between classes.

4. Conclusion

This paper has provided an exhaustive review of the use of deep learning technologies in facial recognition, emphasizing the transformative impact of advanced neural network architectures and innovative training methodologies. Starting from basic neural network frameworks like CNNs and RNNs to more sophisticated architectures such as AlexNet, GoogleNet, VGGNet, and ResNet, this study outlines how these technologies have progressively enhanced the accuracy and efficiency of facial recognition systems. The exploration of data preprocessing techniques—such as face detection, alignment, and normalization—underscores their critical role in standardizing input data, which significantly improves the reliability of subsequent feature extraction processes. Moreover, the paper discusses various feature extraction methodologies that leverage both CNN and GAN frameworks, providing insights into their capability to extract nuanced facial features from complex image data. The assessment of diverse loss functions, including Euclidean distance, angular/cosine margin loss, and several variants of softmax loss, reveals how these methods refine the discriminative power of facial recognition systems. Additionally, the transition from traditional models using eigenfaces and feature descriptors like LBP to modern deep learning models employing deep identity features and complex loss functions such as triplet loss has been thoroughly examined.

Despite substantial advancements, the field of facial recognition still faces several challenges that future research needs to address. One significant area of focus will be enhancing the robustness of facial recognition systems against variations in environmental conditions, such as lighting and occlusions, which currently degrade performance. Future studies could explore more adaptive and resilient neural network architectures that can handle such variability more effectively. Another promising research direction involves improving the training efficiency and reducing the computational demands of deep learning models, making them more accessible for real-time applications on devices with limited processing capabilities. Additionally, there is a growing need to develop privacy-preserving facial recognition technologies that can offer reliable identification while protecting individuals' data from misuse. This could involve innovations in encrypted computation and federated learning approaches. Lastly, the ethical implications of facial recognition technology, such as biases in algorithmic decision-making, necessitate thorough investigation to ensure fairness and equity in automated facial recognition across diverse populations.

References

[1] O'Toole A J, Roark D A, Abdi H 2002 Recognizing moving faces: A psychological and neural synthesis Trends in Cognitive Sciences 6(6) 261–266

DOI: 10.54254/2755-2721/110/2024MELB0107

[2] Dantcheva A, Chen C, Ross A 2012 Can facial cosmetics affect the matching accuracy of face recognition systems? 2012 IEEE Fifth International Conference on Biometrics: Theory, Applications and Systems (BTAS) 391–398 IEEE
[3] Guo G, Zhang N 2019 A survey on deep learning based face recognition Computer Vision and Image Understanding 189 102805
[4] Scherer D, Müller A, Behnke S 2010 Evaluation of pooling operations in convolutional architectures for object recognition International Conference on Artificial Neural Networks 92–101 Berlin, Heidelberg: Springer Berlin Heidelberg
[5] Coşkun M, Uçar A, Yildirim Ö, Demir Y 2017 Face recognition based on convolutional neural network International Conference on Modern Electrical and Energy Systems (MEES) 376–379 IEEE
[6] Abbaspour S, Fotouhi F, Sedaghatbaf A, Fotouhi H, Vahabi M, Linden M 2020 A Comparative Study of Hybrid Deep Learning Models for the Recognition of Human Activity Sensors 20(19) doi: 10.3390/s20195707
[7] Fang W, Chen Y, Xue Q 2021 An overview of the literature on spatiotemporal sequence prediction algorithms based on RNNs Journal on Big Data 3(3) 97
[8] Hochreiter S, Schmidhuber J 1997 Long short-term memory Neural Computation 9(8) 1735–1780
[9] Graves A 2013 Using recurrent neural networks to generate sequences arXiv preprint arXiv:1308.0850
[10] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Rabinovich A 2015 Going deeper with convolutions Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 1–9
[11] He K, Zhang X, Ren S, Sun J 2016 Deep residual learning for image recognition Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 770–778
[12] Simonyan K, Zisserman A 2014 Very deep convolutional networks for large-scale image recognition arXiv preprint arXiv:1409.1556
[13] Schroff F, Kalenichenko D, Philbin J 2015 Imagenet classification with deep convolutional neural networks Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 815–823
[14] Ranjan R, Castillo C D, Chellappa R 2017 L2-constrained softmax loss for discriminative face verification arXiv preprint arXiv:1703.09507
[15] Turk M, Pentland A 1991 Eigenfaces for recognition Journal of Cognitive Neuroscience 3(1) 71–86
[16] Sid Ahmed S, Messali Z, Ouahabi A, Trepout S, Messaoudi C, Marco S 2015 Methods of nonparametric denoising utilizing sharp frequency localization and contourlet transform: Use with electron microscope pictures with a short exposure time Entropy 17(5) 3461–3478
[17] Sirovich L, Kirby M 1987 Low-dimensional procedure for the characterization of human faces JOSA A 4(3) 519–524
[18] Kirby M, Sirovich L 1990 Utilizing the Karhunen-Loeve method to characterize human faces IEEE Transactions on Pattern Analysis and Machine Intelligence 12(1) 103–108
[19] Ahonen T 2004 Face recognition with local binary patterns ECCV