IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 42, No. 1, February, 2025, pp. 353-362
Journal homepage: http://iieta.org/journals/ts

Residential Access Control Facial Recognition Based on Improved ResNet Computing Model

Nan Deng[1,2], Jianfang Lin[3*]

[1] School of Mathematics and Computer Science, Hebei Minzu Normal University, Chengde 067000, China
[2] The Technology Innovation Center of Cultural Tourism Big Data of Hebei Province, Chengde 067000, China
[3] Department of Computer Science and Technology, Tangshan Normal University, Tangshan 063000, China

Corresponding Author Email: 131051@tstc.edu.cn

https://doi.org/10.18280/ts.420130

Received: 18 July 2024
Revised: 22 November 2024
Accepted: 3 December 2024
Available online: 28 February 2025

Keywords:
ResNet, residential access control, facial recognition, identity mapping, loss function

ABSTRACT

Dynamic facial recognition is extremely crucial in residential access control systems, but factors such as multiple poses, facial expressions, and side faces often affect its accuracy. To address this challenge, research has proposed an improved Residual Network (ResNet) computing model. This method first introduces Convolutional Neural Network s (CNNs) and ResNet models, then improves the loss function by introducing a joint loss, and finally constructs the ResNet identity mapping model. The research results demonstrate that the facial verification accuracy of this model on the Labeled Faces in the Wild (LFW), YouTube Faces (YTF), and Similar-looking Labeled Faces in the Wild (SLLFW) datasets reaches 99.6% and 96%, respectively, surpassing the other three loss function models. In the M-Face and F-Scrub datasets, the model exhibits excellent recognition and validation performance. On the LFW and YTF datasets, the facial recognition accuracy is 99.6% and 95.7%, respectively, while on the Intelligence Advanced Research Projects Activity (IARPA) Janus Benchmark-B (IJBB) and IARPA Janus Benchmark-C (IJBC) datasets, it achieves scores of 0.922 and 0.945. In video face recognition analysis, the model remains effective even when the face is significantly deviated. Real-time performance and frame per second transmission tests show that the system operates at 0.39 times the speed of the original video, with an average frame rate of 27-29 frames per second. The improved ResNet model proposed in the study performs well in handling multi-pose and side face recognition, which is of great significance for the improvement of facial recognition technology in residential access control systems.

1. INTRODUCTION

With the rapid development of computer science and artificial intelligence, people's demand for life safety and convenience is increasing day by day. Under this trend, residential access control systems have received widespread attention as an important component of safe and convenient living. Facial recognition technology, known for its non-contact and high efficiency, has become an ideal choice for improving the performance of access control systems. However, current static facial recognition technology faces many challenges in practical applications, especially in terms of recognition accuracy in dynamic environments [1-3].

The main research challenge focuses on facial recognition in dynamic environments, such as multiple poses, changes in facial expressions, and recognition performance under different lighting conditions. These factors greatly increase the complexity of facial recognition in access control systems, affecting the accuracy and stability of recognition. For example, changes in profile or facial expressions may lead to recognition failure, reducing system efficiency and user experience. Additionally, recognition under different lighting conditions is also a technical challenge, as changes in light can significantly affect recognition results [4-6].

To address these challenges, a residential access control facial recognition method based on an improved ResNet computing model is proposed. This method initially introduces CNNs and ResNet, then enhances the loss function and incorporates joint loss to improve the model's performance in handling multiple poses and side face recognition. Finally, by refining the identity mapping of ResNet, the model's ability to recognize faces in complex environments is enhanced. The overall structure of the study includes four parts. The first part summarizes the relevant research achievements and shortcomings of neural networks and face recognition both domestically and internationally. In the second part, the study begins with the proposal of CNNs and ResNet, then goes on to improve the loss function and introduces joint losses based on this foundation. Subsequently, ResNet is further refined, and a ResNet identity mapping model is successfully constructed.

In the third part, the research experiment compares and analyzes the proposed improved ResNet. The fourth part summarizes the experimental results, highlights the shortcomings of the research, and proposes future research directions.

2. RELATED WORKS

With the rapid advancement of big data and computing power, deep learning technology has made significant strides in the field of facial recognition, substantially enhancing both the accuracy and efficiency of recognition. Concurrently, as awareness of privacy protection strengthens, research is increasingly focused on improving recognition performance while safeguarding personal privacy. These developments have spurred innovation in algorithms and deepened our understanding of the practical application of deep learning in real-world scenarios [7]. Below, we introduce some related research by scientists and scholars.

Handa et al. [8] proposed a model based on Deep Convolutional Neural Networks (DCNN) to address the challenges of Facial Expression Recognition (FER). The model begins by selecting an appropriate activation function based on its accuracy and training loss over a database. It then employs an incremental strategy, developing deeper models simultaneously from shallower networks to increase accuracy with less training loss. By employing an ensemble of CNNs and DCNNs, the model achieves an accuracy of 74.15% on the Facial Expression Recognition 2013 (FER2013) dataset, 96.20% on the Extended Cohn-Kanade (CK+) dataset, and 98.25% on the Japanese Female Facial Expression (JAFFE) dataset, outperforming previous works.

Yang and Zhang [9] utilized a multi-task cascaded convolutional neural network (MTCNN) for heterogeneous face feature detection. The algorithm takes full advantage of an image pyramid, boundary regression, fully convolutional attention networks, and non-maximum suppression. The approach begins with a candidate frame plus classifier for fast and efficient face detection. The candidate window is generated by the proposal network (P-Net), with the high-precision candidate window filtered and selected by the refinement network (R-Net), and the final bounding box and facial keypoints generated by the output network (O-Net). To demonstrate the effectiveness of this method in visible light, near-infrared, and sketch face recognition scenes, it was verified using the datasets of CUHK Face Sketch Database (CUFS), CUHK Face Sketch FERET Database (CUFSF), and the Chinese Academy of Sciences Institute of Automation Near-Infrared and Visible light (NIR-VIS) 2.0. Experiments show that this method is effective for heterogeneous face images and surpasses the latest algorithms.

Tsai and Chi [10] proposed a multi-task training method based on a feature pyramid and triplet loss to train a single-stage face detection and face recognition deep neural network. This method utilizes the same backbone network for data transfer, achieving weight sharing and reducing computational complexity. Feature pyramids are used to locate faces, and feature matching is performed through simple functions. The network is trained with a triplet loss function to learn a discriminative feature representation, achieving state-of-the-art performance in both face detection and recognition tasks.

Abed et al. [11] proposed a new method for extracting keyframes from videos based on face quality and deep learning for a face recognition task. The initial step involves face detection using the MTCNN detector, which identifies five landmarks (the eyes, the two corners of the mouth, and the nose), limits face boundaries within a bounding box, and provides a confidence score. The first phase aims to generate the face quality score for each face in the dataset prepared for the learning step, using three face feature extractors: Gabor Filter (Gabor), Local Binary Pattern (LBP), and Histogram of Oriented Gradients (HoG). The second phase involves training a DCNN in a supervised manner to select frames with the best face quality. The results demonstrate the effectiveness of the proposed method compared to state-of-the-art methods.

Yallamandaiah and Purnachand [12] proposed a method that combines a guided image filter and a CNN. The guided image filter, a smoothing operator that performs well near edges, is used to enhance the quality of facial images. This enhancement significantly improves the performance of the CNN used for facial feature extraction and recognition. The method is specifically designed to address challenges faced by face recognition systems in unconstrained environments, where factors such as varying lighting conditions, facial expressions, and occlusions can significantly affect recognition accuracy. The combination of the guided image filter and CNN has proven effective in improving recognition rates.

Putra et al. [13] introduced a webinar student presence system using face recognition with a Regional Convolutional Neural Network (R-CNN). The system, aimed at accurately tracking student attendance during webinars, utilizes object detection across several scenarios. This research focuses on designing an accurate student presence system using the R-CNN method, particularly under conditions of sufficient lighting, which is crucial for the face recognition process. The system is designed to be efficient and effective in various lighting conditions and with different facial expressions.

Zhang et al. [14] presented a model in the "International Journal of Biometrics" that addresses the challenge of age interference in deep learning-based face recognition. Their approach focuses on developing a fast and accurate face recognition system capable of effectively handling variations in facial features due to age. The model incorporates advanced deep learning techniques to minimize the impact of age on recognition accuracy, ensuring that the system can quickly and reliably identify individuals despite changes in their appearance over time.

Lin et al. [15] proposed the deep representation alignment network (DRA-Net), a representation alignment framework that includes a denoising autoencoder (DAE) and an innovative deep representation transformation (DRT) block to learn identity-preserving representations. This work achieves state-of-the-art performance on the LFW, YTF, Multi Pose, Illumination, and Expressions (Multi-PIE), China Family Panel Studies (CFP), IARPA Janus Benchmark-A (IJB-A), and Multi-Yaw Multi-Pitch High-Quality Database for Facial Pose Analysis (M2FPA) datasets.

In summary, while many experts and scholars have made significant progress in improving facial recognition technology and applying it to multiple fields, the issues of multi-pose and side-face recognition in dynamic environments remain challenging. Dynamic face detection and recognition face more complex problems than static face recognition due to the diversity of poses and expressions. To address this, a computational model for CNNs and ResNet was proposed, incorporating a joint loss function to further enhance ResNet and construct an identity mapping model. This model has performed well in handling multi-pose and side-face recognition, marking a significant advancement in improving dynamic facial recognition technology.

3. ENHANCED RESIDENTIAL ACCESS CONTROL USING AN ADVANCED RESNET FACIAL RECOGNITION MODEL

This study introduces a computational model utilizing CNNs and ResNet, enhancing the loss function through the incorporation of joint losses. Given that facial recognition is frequently compromised by variations in pose and facial expressions, the ResNet model has been further refined, culminating in the successful development of a ResNet identity mapping model.

3.1 Construction of computational models for CNNs and ResNet

The objective of this research is to adapt the ResNet computing model for use in facial recognition within residential access control systems. Dynamic facial recognition is critical in these systems, as it must efficiently process the continually changing facial expressions and angles captured in video streams. The deployment of deep learning techniques, particularly those founded on ResNet models, effectively meets these challenges.

The implementation of the system unfolds through several essential steps. Initially, facial detection and alignment are conducted to accurately extract facial images from video streams. Subsequently, feature extraction takes place. Regardless of the technology employed, the precise capture of facial features is imperative for the subsequent recognition phase. The process concludes with the confirmation of identity by comparing the similarity of facial features across various images. This comprehensive procedure is illustrated in Figure 1.

The ongoing enhancements to CNN architecture have significantly advanced various recognition tasks. The structure of a CNN is reminiscent of the human visual system, making it exceptionally capable of processing two-dimensional and three-dimensional images, and adept at learning and extracting features [16, 17]. A CNN comprises a feature extractor and a classifier. The input data is processed layer by layer, through a series of convolutional and max pooling layers. The convolutional layer is tasked with feature extraction, while the max pooling layer reduces the dimensions of these features. Features are transmitted between network layers, and output feature maps are generated through either linear or nonlinear activation functions, as depicted in Eq. (1).

$$x_j^l = f\left(\sum_{i \in M_j} x_i^{l-1} * k_{ij}^l + b_j^l\right) \tag{1}$$

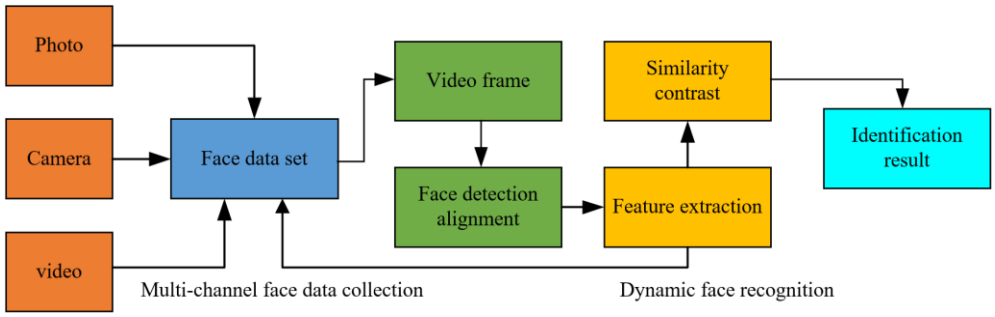


Figure 1. Residential access control system dynamic face recognition

where, x_j^l represents the input, x_i^{l-1} represents the output of the previous layer, and b_j^l represents the bias term. The pooling layer is used to reduce the size of the input feature map, thereby reducing the computational burden on subsequent layers of the network. By performing scanning operations on the input feature map, the pooling layer calculates the size reduction operation for each window, which may include taking the maximum value within the region or average pooling. These pooling operations help extract important features while reducing data dimensions, as shown in Eq. (2).

$$\begin{cases} Z_j^l = \max\{y_i^{l-1}\} = \max_{i \in K_i} y_i^{l-1} \\ Z_j^l = mean\{y_i^{l-1}\} = \dfrac{1}{|K_j|}\sum_{i \in K_j} y_i^{l-1} \end{cases} \tag{2}$$

where, K_i represents the selected ith region. This area covers the jth neuron in the first layer and its corresponding neuron set y_i^{l-1} in the previous layer, which is the same size as the pooled nucleus. The backend of a network structure typically consists of fully connected layers, which convert extracted features into outputs for classification or regression tasks. To improve efficiency and performance, these dense layers are often replaced by average pooling layers, especially in parameter dense networks. In addition, the contrastive loss function plays a crucial role in complex tasks such as facial recognition, as it improves the recognition ability of the model by distinguishing positive and negative sample pairs. The specific formula is detailed in Eq. (3).

$$\begin{aligned} L = {} & y_{ij} \max\left(0, \| f(x_i) - f(x_j) \|_2 - {}^{+}\right) \\ & + \left(1 - y_{ij}\right) \max\left(0, {}^{-} - \| f(x_i) - f(x_j) \|_2\right) \end{aligned} \tag{3}$$

The most commonly used loss function currently is Softmax loss, which is widely popular due to its simplicity and no need for additional hyperparameters. The loss function maps the extracted features to the range of 0 to 1 and performs recognition classification based on the obtained probability, as shown in Eq. (4).

$$L_s = -\frac{1}{N}\sum_{i=1}^{N} \log\left(\frac{e^{f_{it}}}{\sum_{j=1}^{C} e^{f_j}}\right) \tag{4}$$

where, C represents the total number of different categories,

while N refers to the sample size for each batch of processing. Due to the limitations of Softmax in dealing with uneven distribution within classes and distinguishing similar samples, center loss is introduced to reduce intra class differences. By combining Softmax loss and center loss, samples can be trained simultaneously to achieve intra-class compactness and inter class dispersion, as shown in Eq. (5).

$$\begin{cases} L_C = \frac{1}{2}\sum_{i=1}^{N} \| x_i - Cy_i \|_2^2 \\ L = L_s + \lambda L_c = -\frac{1}{N}\sum_{i=}^{N} \log\left(\frac{e^{f_{yi}}}{\sum_{j=1}^{C} e^{f_j}} \right) + \frac{\lambda}{2}\sum_{i=1}^{N} \| x_i - C_{y_u} \|_2^2 \end{cases} \tag{5}$$

where, the center point C_{y_i} of each category is adjusted with the update of the feature x_i of each sample, to reduce the distance between samples within the class.

3.2 Construction of joint loss function model for ResNet

To adapt to dynamic facial recognition environments, a fusion loss function is further proposed, which combines Softmax loss, center loss, and marginal loss to reduce intra class differences and increase inter class separation. Improvements have been made on the ResNet34 architecture, such as reducing the size of the convolutional kernel to extract more facial information, and replacing the original activation function. In addition, ResNet identity mapping layer has been added to improve the recognition rate of lateral faces. In multi class classification, Softmax maps multiple outputs to the range of 0 to 1, as shown in Eq. (6).

$$S_i = \frac{e^{V_i}}{\sum_j e^{V_j}} \tag{6}$$

The Softmax layer structure process can be represented as shown in Figure 2.

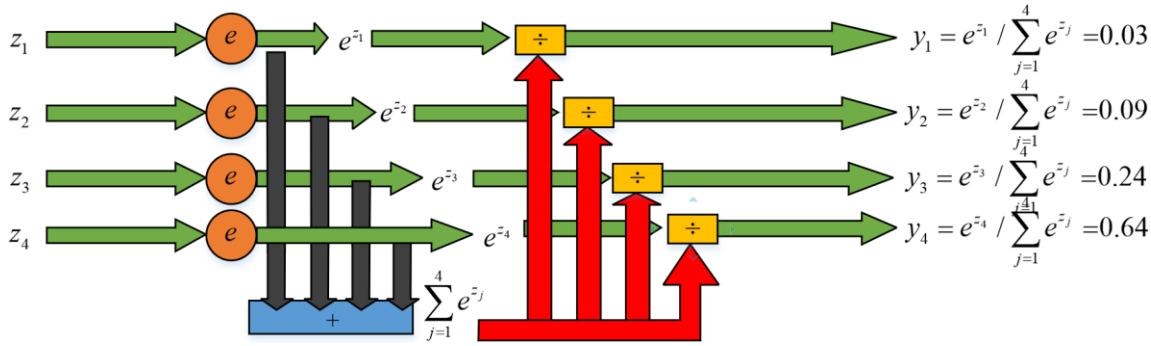


Figure 2. Softmax structure

Figure 2 illustrates the Softmax operation, which converts the initial response of the network into a distribution probability. To fully understand Softmax loss, it is necessary to first understand the cross entropy function, which is a combination of Softmax function and cross entropy. Cross entropy evaluates the difference between the actual probability output q and the predicted probability output p, with lower values indicating higher similarity between the two [18]. The definition of cross entropy $H(p,q)$ for variable x is shown in Eq. (7).

$$H(p,q) = \sum_x p(x) \log \frac{1}{q(x)} = -\sum_x p(x) \log q(x) \tag{7}$$

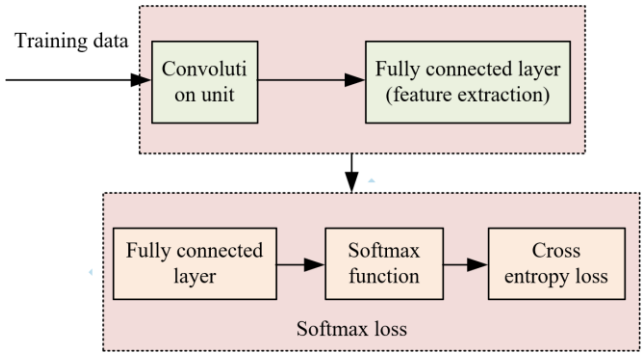


Figure 3. Softmax loss classification flow chart

When performing image classification in CNNs, the training samples first pass through the network and output feature maps through fully connected layers. This feature map is multiplied by the weight matrix $W = \{W_1, W_2, W_3, \ldots, W_{K-1}, W_K\}$ of the classification layer to generate scores for each category. Next, these scores are processed using the Softmax function to generate normalized category probabilities, and the cross entropy loss function is calculated. Softmax is particularly skilled at amplifying small differences between categories, which is crucial in the optimization process. The specific workflow is shown in Figure 3.

In addition to precise classification of known categories, evaluating similarity is essential. Merely classifying known species accurately does not suffice to enhance generalization. It is crucial to ensure consistency within samples of the same type and discrimination between different types. However, these aspects are not directly optimized by Softmax loss. To address this, a new joint loss strategy combining center loss with Softmax loss has been developed, as detailed in Eq. (8).

$$\begin{cases} L_C = \frac{1}{2}\sum_{i=1}^{N} \| f_i - c_{yi} \|_2^2 \\ L = L_S + \lambda L_C = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{e^{W_{yi}^T f_i + b_{yi}}}{\sum_{j=1}^{K} e^{W_j^T f_i + b_j}} + \frac{\lambda}{2}\sum_{i=1}^{N} \| f_i - c_{yi} \|_2^2 \end{cases} \tag{8}$$

where, C_{yi} represents the center point of class yi, while λ is a hyperparameter that balances different loss types. After combining Softmax loss and center loss, intra-class consistency has been significantly improved. However, relying solely on Softmax loss to constrain inter class differences is insufficient, as it mainly promotes feature differentiation. Therefore, the study introduced a joint supervision method of marginal loss and Softmax loss, as shown in Eq. (9).

$$\begin{cases} L = L_S + \lambda L_{Mar} \\ L_{Mar} = \dfrac{1}{N^2 - N} \sum\limits_{i,j,i \neq j}^{N} \left(\xi - y_{ij} \left(\left\| \dfrac{f_i}{\| f_i \|} - \dfrac{f_j}{\| f_j \|} \right\|_2^2 \right) \right) \end{cases} \tag{9}$$

where, f_i and f_j represent the vectors of the ith and jth in the batch. ξ represents the fault tolerance range beyond the classification decision boundary. Next, a fusion loss function is proposed based on the concepts of Softmax loss, center loss, and marginal loss. Softmax is responsible for classification accuracy, while joint loss adjusts inter class distance. It sets a standard for joint loss using the class centers determined by the center loss, and introduces penalty terms for these pairs to increase the total loss, as shown in Eq. (10).

$$\begin{cases} L = L_S + aL_C + \beta L_J \\ L_J = \sum\limits_{i,j=1}^{K} \max\left(\left\| c_i - c_j \right\|_2^2 - M, 0 \right) \end{cases} \tag{10}$$

where, a and β are the key parameters for adjusting the center loss and joint loss, respectively. c_i and c_j represent the core points of i and j classes, while M is the set minimum distance threshold.

3.3 Identity mapping model based on improved ResNet

Due to the influence of multiple poses and facial expressions on facial recognition, it is necessary to improve ResNet34 by removing the max pooling layer after the first convolutional layer to preserve more localization information. To stabilize the output distribution and accelerate the learning process, a batch normalization layer is added at the entrance of the residual unit, as shown in Figure 4. Batch normalization maintains the excitation values in the sensitive area of nonlinear functions by standardizing input values, avoiding gradient vanishing, and improving training speed.

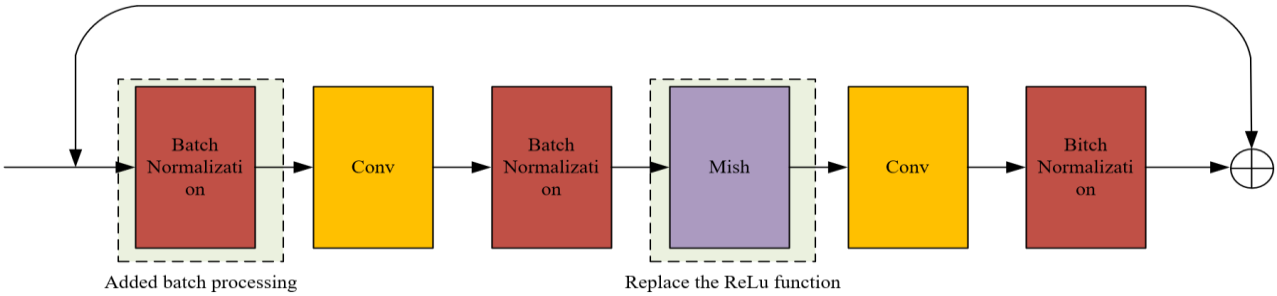


Figure 4. Improvement of residual module

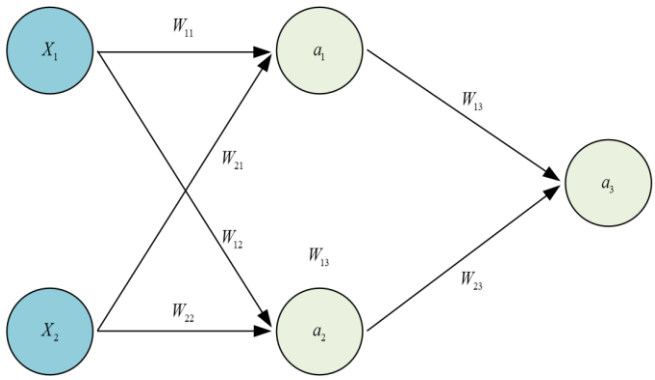


Figure 5. Neural network structure diagram

Before building a deep learning network, it is essential to initialize the weight parameters, W, of the neural network correctly. Training these models involves iterative updates to W. As the number of layers in the network increases, the risk of encountering vanishing or exploding gradients during the optimization process also rises, making proper weight initialization critical. While weights W and bias b in linear and logistic regression are typically initialized to zero, this approach can lead to failure in neural networks, particularly when hidden layers are involved, as noted in references [19, 20].

Figure 5 underscores the importance of proper weight initialization, illustrating why weights must not be set to zero. Moreover, traditional gradient descent optimization algorithms often fall short with large-scale datasets. In response, the Adam algorithm is introduced, which integrates the benefits of the momentum method and the RMS Prop algorithm to provide rapid and efficient optimization. Each iteration of Adam uses the exponential weighted moving average of the squared small batch random gradient in the momentum variable V and RMS Prop, as detailed in Eq. (11).

$$\begin{cases} V_t \leftarrow \beta_1 V_{t-1} + (1 - \beta_1) g_t \\ S_t \leftarrow \beta_2 S_{t-1} + (1 - \beta_2) g_t \odot g_t \end{cases} \tag{11}$$

Then bias correction on variables V and S is performed, as shown in Eq. (12).

$$\begin{cases} \hat{V}_t \leftarrow \dfrac{V_t}{1 - \beta_1^t} \\ \hat{S}_t \leftarrow \dfrac{S_2}{1 - \beta_2^t} \\ g_t' \leftarrow \dfrac{\eta \hat{V}_t}{\sqrt{\hat{S}_t + \in}} \end{cases} \tag{12}$$

where, η represents the initial learning rate greater than zero,

but ∈ is a fixed constant. To avoid overfitting during training, normalization is applied to the loss function by introducing parameters representing model complexity, as shown in Eq. (13).

$$\begin{cases} J = J_0 + \lambda \sum_w |w| \\ J = J_0 + \frac{\lambda}{2} \sum_w w^2 \end{cases} \tag{13}$$

It is difficult to learn deep facial representations with geometric stability under significant pose changes. Assuming there is a mapping relationship between the sides and the front, the differences between them in the feature space can be connected through consistency mapping, and it is possible to obtain the mapped front features through the side features, thereby improving the recognition ability of the sides. It sets the positive variable as x_f and the lateral variable as x_p, treat CNN as a function, and the goal is to find the mapping M_g. The residual model is used to obtain Eq. (14).

$$\begin{aligned} \phi(gX_p) &= M_g \phi(X_p) \\ &= \phi(x_p) + \gamma(X_p) R(\phi(X_p) \approx (X_f) \end{aligned} \tag{14}$$

At the basic feature output end of the network, a constant residual mapping section is introduced to transform the initial feature representation to achieve the final feature representation, as shown in Figure 6.

It can be observed that two red boxes configured with Mish activation functions represent fully connected layers trained independently of the main network. The training objective of this independent branch is to reduce the Euclidean distance between the side and the corresponding front, as shown in Eq. (15).

$$\min_{\Theta_R} E \| \phi(x) + y(x) + \Re(\phi(x)); \Theta_R - \phi(x_f) \|_2^2 \tag{15}$$

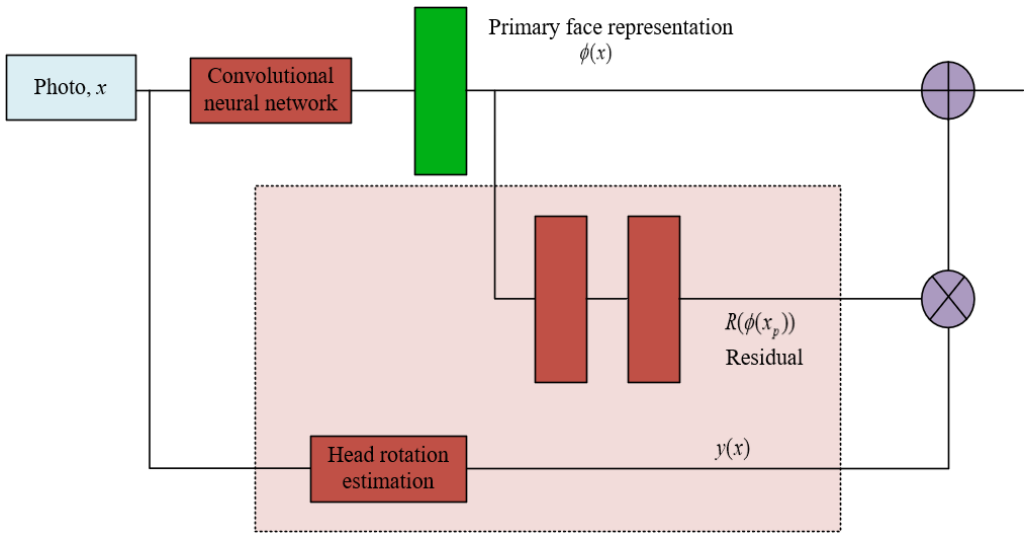


Figure 6. ResNet identity mapping example diagram

The head rotation estimator is used to calculate the offset angle $y(x)$ between the image and the front. Firstly, key points are extracted through facial alignment, and then the EPNP algorithm is used to estimate the head rotation angle. The rotation angle is mapped to the range of [0,1] through the function $\sigma\left(\frac{4}{\pi}y - 1\right)$. The training method is to combine the trained identity mapping module with the main network, and conduct end-to-end overall training, specifically training the identity mapping module after end-to-end training.

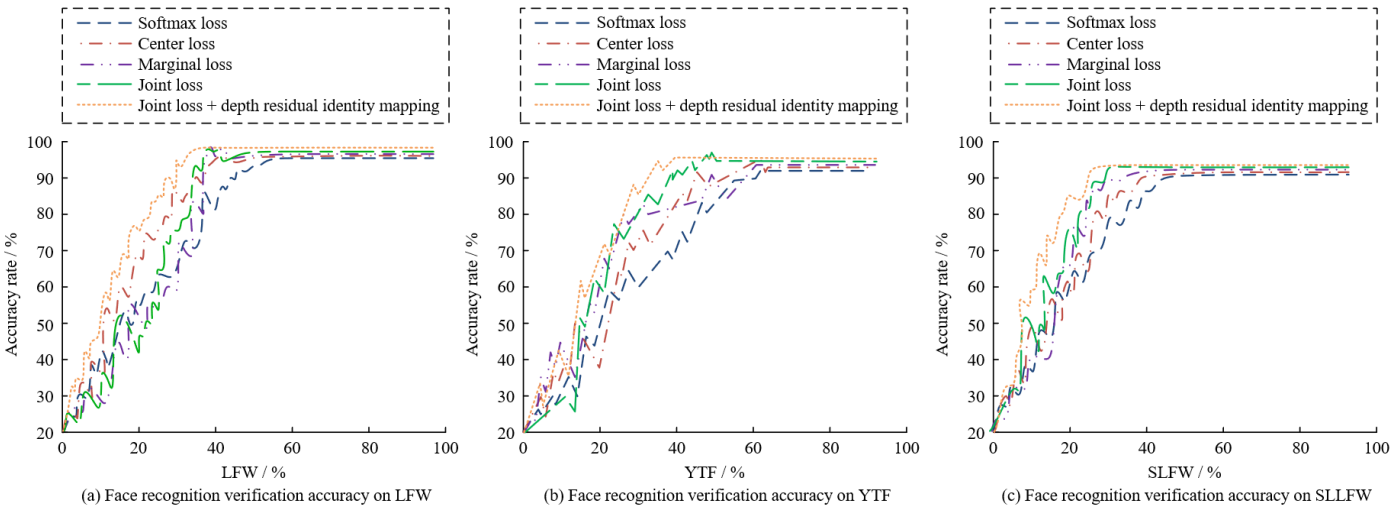
4. ANALYSIS OF FACIAL RECOGNITION RESULTS FOR RESIDENTIAL ACCESS CONTROL USING AN IMPROVED RESNET COMPUTING MODEL

This study began by introducing five distinct datasets and evaluating the effectiveness of the improved ResNet model through rigorous testing on these datasets. Following this, the enhanced model was benchmarked against other contemporary facial recognition models. Finally, the effectiveness of the improved ResNet model for dynamic facial recognition in video was validated, demonstrating its robust capabilities.

4.1 Result analysis of loss function and identity mapping module based on improved ResNet

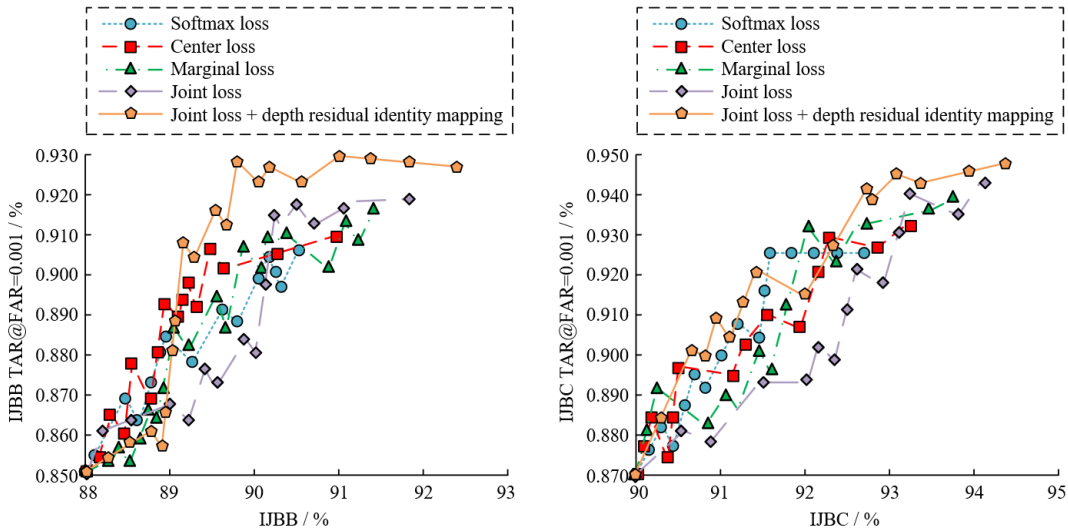
The focus of this study was on facial recognition for residential access control using an enhanced ResNet-34 network model. The experiments compared five different facial recognition models employing various loss functions: Softmax loss, center loss, marginal loss, and joint loss function, as well as a hybrid model combining joint loss with ResNet identity mapping. The selected datasets for this study included M-Face, SLLFW, LFW, YTF, F-Slub, IJBB, and IJBC, with hyperparameters set to α=8, β=6, and M=250.

The LFW dataset, a widely recognized benchmark in facial recognition, was used to randomly select 5500 pairs of face images, comprising 2750 pairs each of positive and negative samples. The YTF dataset consists of approximately 3400 videos sourced from YouTube, averaging about two videos per subject. The SLLFW dataset, testing the same frontal face pairs as LFW, increased the difficulty by substituting the random negative samples in LFW with 2800 to 3100 pairs of similar-looking faces. The comparative results of these five models across the three datasets—LFW, YTF, and SLLFW—are illustrated in Figure 7.



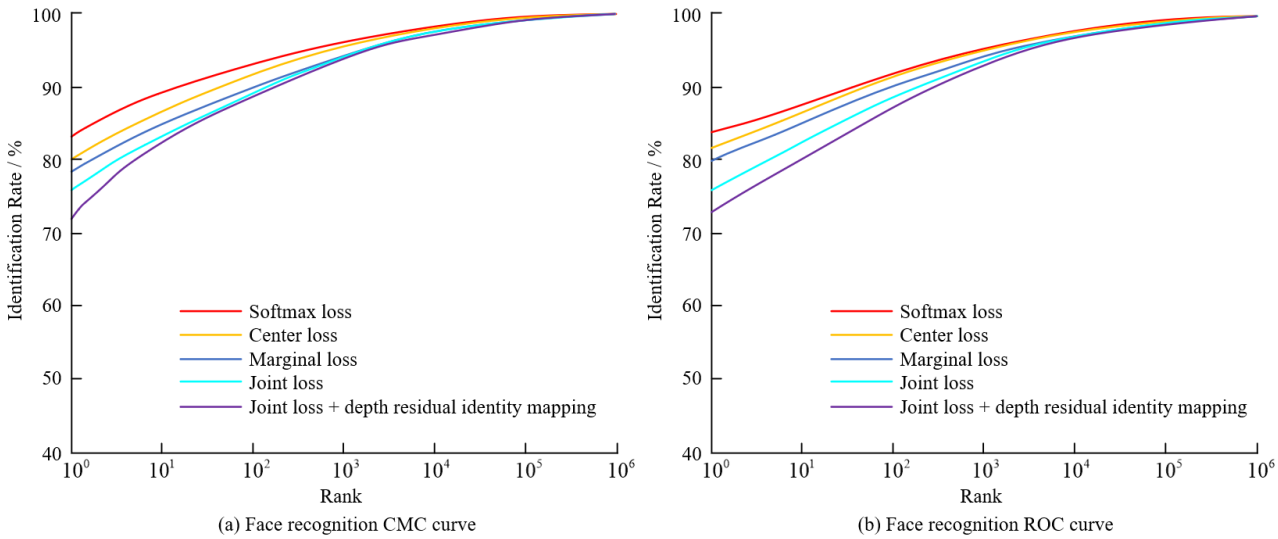
(a) Face recognition verification accuracy on LFW (b) Face recognition verification accuracy on YTF (c) Face recognition verification accuracy on SLLFW

Figure 7. Face recognition verification accuracy of LFW, YTF and SLLFW datasets



(a) Face recognition verification accuracy of IJBB data set (b) Face recognition verification accuracy of IJBC data set

Figure 8. Face recognition verification accuracy of IJBB and IJBC datasets



(a) Face recognition CMC curve (b) Face recognition ROC curve

Figure 9. Face recognition CMC and ROC curve results

Figure 7 demonstrates that the model combining joint loss and the ResNet identity mapping module achieves the highest facial verification accuracy on two datasets, reaching 99.6% and 96%, respectively. Specifically, this model shows a 1.5% improvement over traditional Softmax loss in Figure 7(a), surpasses Center loss by 1.3%, exceeds Marginal loss by 1.05%, and is 0.8% higher than using joint loss alone. In Figure 7(b), the same model records a 0.9% improvement over

Softmax, 0.35% over Center loss, 0.3% over Marginal loss, and 1.0% over joint loss alone. Although Figure 7(c) indicates a decrease in accuracy on the SLLFW dataset, the model incorporating joint loss and ResNet identity mapping still outperforms other models. On SLLFW, this model is 1.2% better than Softmax, 0.4% better than Center loss, and 0.25% better than joint loss alone. The comparisons of the five models on the IJBB and IJBC datasets are presented in Figure 8. The IJBB dataset consists of approximately 56,000 frames and 22,000 static images from around 7,000 videos, with about 11,000 true matches and 80,000 non-matches. IJBC, a larger version of IJBB, contains 32,000 static images. The model combining joint loss and ResNet identity mapping displays significant performance advantages on both datasets.

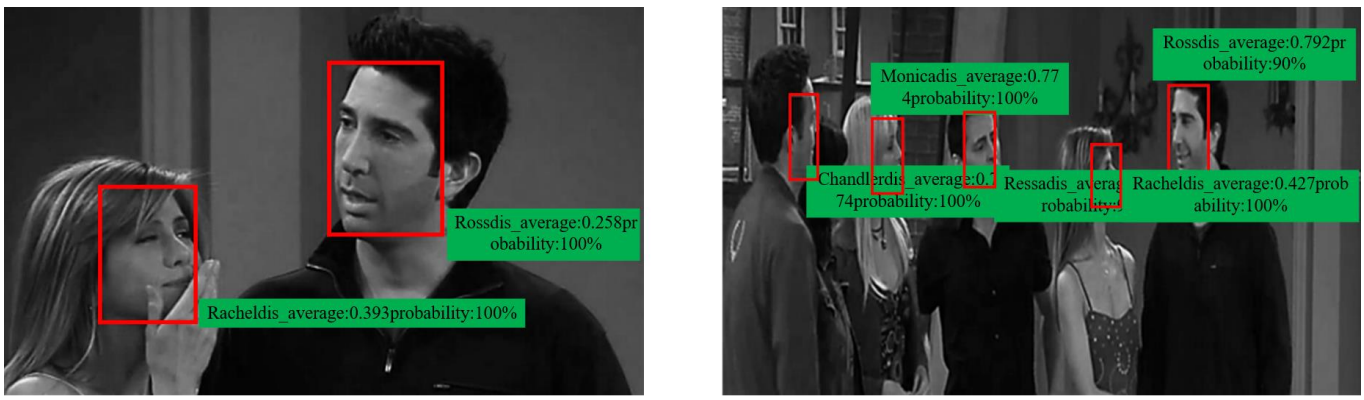
Next, experiments would be conducted using the M-Face and F-Scrub datasets, as shown in Figure 9. The M-Face dataset contained 1000000 face images for interference testing. The F-Scrub dataset contained 100000 images for testing. The study used cumulative matching characteristic curves and receiver operation characteristic curves for evaluation. The cumulative matching characteristic curve was used to measure the performance of different models in recognition and validation testing, while the receiver operation characteristic curve was used to evaluate the performance of different models in 1:1 matching. The experimental results showed that the model combining joint loss and ResNet identity mapping performed better in recognition and validation testing, proving the effectiveness of the model.

4.2 Analysis of residential access control video facial recognition results based on improved ResNet

The previous experiment demonstrated that the improved ResNet had good performance in facial recognition. By incorporating joint loss and deep residual identity mapping layers into the improved ResNet34 network, it performed well. To further demonstrate the effectiveness of the model, experimental comparisons were conducted with other existing models, and the results are shown in Table 1. The algorithm proposed in the study had facial recognition accuracy of 99.6% and 95.7% on the LFW and YTF datasets, respectively. The facial recognition accuracy on the IJBB and IJBC datasets was as high as 0.922 and 0.945, respectively. Compared with other models, the model proposed in this study still had certain advantages, further demonstrating the effectiveness of the model.

Table 1. Four data sets in different methods of face recognition verification accuracy

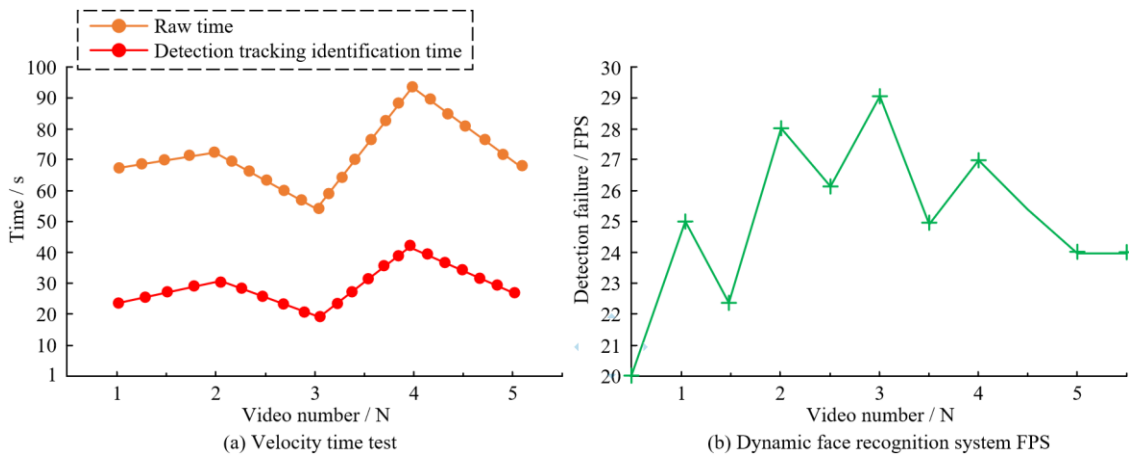
Model	LFW Data Set (%)	YTF Data Set (%)	IJBB TAR@FAR=0.001	IJBC TAR@FAR=0.001
VGG-Face	99.0	94.4	0.910	0.930
Deep-Face	97.3	91.3	0.900	0.915
DeepID	99.4	93.0	0.905	0.920
Study the proposed model	99.6	95.7	0.923	0.950



(a) Video face recognition effect 1

(b) Video face recognition effect 2

Figure 10. Dynamic face recognition rendering



(a) Velocity time test

(b) Dynamic face recognition system FPS

Figure 11. System speed time and FPS

The next phase of this study involved conducting video facial recognition analysis. Initially, it was necessary to preprocess TV programs from the Kaggle dataset and store the feature vectors of various individuals in a database, with each individual represented by 50 images, covering a total of 100 individuals. The study utilized MTCNN for face detection and Simple Online and Realtime Tracking with a Deep Association Metric (DeepSort) for tracking. Faces detected in the video were then compared with feature vectors in the database. If a face matched at least 45 feature vectors, the system would output labels and probabilities, utilizing this model for continuous face tracking as illustrated in Figure 10.

As depicted in Figure 10, the model successfully identified faces in the video and matched them against the database entries. The outputs included human labels, Euclidean distances, and probabilities, indicating that the enhanced facial recognition model could effectively recognize characters in videos, even with significant facial deviations, thereby confirming the model's effectiveness.

To assess the real-time performance and Frames Per Second (FPS) and ensure they met practical requirements, performance testing was conducted, as shown in Figure 11. The results confirmed that the system could operate effectively in real-world scenarios, with processing speeds reaching 0.39 times that of the original video and maintaining an average FPS of 27-29. These results convincingly demonstrated the model's robust capability for dynamic recognition in applications such as residential access control.

5. CONCLUSION

A computational model combining CNNs and ResNet was proposed to address the impact of multiple poses and lighting changes on facial recognition technology in residential access control systems. Based on this, the loss function was further improved by introducing joint losses. This improvement led to the construction of the ResNet identity mapping model. The data showed that based on the improved ResNet-34 network model, the proposed model performed well on different datasets. Especially on the LFW, YTF, and SLLFW datasets, the proposed model achieved facial verification accuracy of 99.6% and 96%, respectively. This performance was superior to models that use traditional Softmax loss, center loss, marginal loss, and individual joint loss. In addition, in the experiments on the M-Face and F-Scrub datasets, the model also performed well in recognition and validation testing. In addition, compared with other models, the improved ResNet34 network achieved face recognition accuracy of 99.6% and 95.7% on the LFW and YTF datasets, respectively, and 0.922 and 0.945 on the IJBB and IJBC datasets, demonstrating its significant advantages. In video face recognition analysis, the model demonstrated strong performance in identifying characters in videos, even in situations where the face was heavily skewed. The real-time performance and FPS test results showed that the system met practical requirements, with a speed of 0.39 times that of the original video and an average FPS of 27-29 FPS. In summary, the improved ResNet model performed well in residential access control facial recognition, especially in handling multi pose and side face recognition. However, there is still room for improvement in the accuracy of the model under extreme lighting and certain deflection angles. Future research can focus on further improving the performance of models under these extreme conditions and exploring more efficient algorithms to reduce computational costs, thereby achieving wider applications.

FUNDINGS

This work is supported by the R&D investment intensity growth reward fund from the Science and Technology Bureau of Chengde City, Hebei Province (Grant No.: 232304B) and Hebei Minzu Normal University (Grant No.: DR2023003).

REFERENCES

[1] Suma, K. (2021). Dense feature based face recognition from surveillance video using convolutional neural network. Turkish Journal of Computer and Mathematics Education, 12(5): 1436-1449.

[2] Ge, Y., Liu, H., Du, J., Li, Z., Wei, Y. (2023). Masked face recognition with convolutional visual self-attention network. Neurocomputing, 518: 496-506. https://doi.org/10.1016/j.neucom.2022.10.025

[3] Shakeel, M.S. (2024). CAAM: A calibrated augmented attention module for masked face recognition. Journal of Visual Communication and Image Representation, 104: 104315. https://doi.org/10.1016/j.jvcir.2024.104315

[4] Bajpai, S., Mishra, G., Jain, R., Jain, D. K., Saini, D., Hussain, A. (2025). RI-L1Approx: A novel Resnet-Inception-based Fast L1-approximation method for face recognition. Neurocomputing, 613: 128708. https://doi.org/10.1016/j.neucom.2024.128708

[5] Lin, D., Wang, H., Lei, X., Min, W., Yao, C., Zhong, Y., Guan, Y. L. (2024). DSU-GAN: A robust frontal face recognition approach based on generative adversarial network. Computer Vision and Image Understanding, 249: 104128. https://doi.org/10.1016/j.cviu.2024.104128

[6] Xing, C., Wang, J.S., Zheng, B.W. (2021). Hybrid face recognition method based on Gabor wavelet transform and VGG convolutional neural network with improved pooling strategy. IAENG International Journal of Computer Science, 48(2): 413-427.

[7] Tamilselvi, M., Karthikeyan, S., Ramkumar, G. (2021). Face recognition based on spatio angular using visual geometric group-19 convolutional neural network. Annals of the Romanian Society for Cell Biology, 25(3): 2131-2138.

[8] Handa, A., Agarwal, R., Kohli, N. (2021). Incremental approach for multi-modal face expression recognition system using deep neural networks. International Journal of Computational Vision and Robotics, 11(1): 1-20. https://doi.org/10.1504/IJCVR.2021.111881

[9] Yang, X., Zhang, W. (2022). Heterogeneous face detection based on multi-task cascaded convolutional neural network. IET Image Processing, 16(1): 207-215. https://doi.org/10.1049/ipr2.12344

[10] Tsai, T.H., Chi, P.T. (2022). A single-stage face detection and face recognition deep neural network based on feature pyramid and triplet loss. IET Image Processing, 16(8): 2148-2156. https://doi.org/10.1049/ipr2.12479

[11] Abed, R., Bahroun, S., Zagrouba, E. (2021). KeyFrame extraction based on face quality measurement and convolutional neural network for efficient face

recognition in videos. Multimedia Tools and Applications, 80(15): 23157-23179. https://doi.org/10.1007/s11042-020-09385-5

[12] Yallamandaiah, S., Purnachand, N. (2021). An effective face recognition method using guided image filter and convolutional neural network. Indonesian Journal of Electrical Engineering and Computer Science, 23(3): 1699-1707. https://doi.org/10.11591/ijeecs.v23.i3.pp1699-1707

[13] Putra, A.T., Usman, K., Saidah, S. (2021). Webinar student presence system based on regional convolutional neural network using face recognition. Jurnal Teknik Informatika, 2(2): 109-118. https://doi.org/10.20884/1.jutif.2021.2.2.82

[14] Zhang, Y., Wu, P., Zhao, J., Feng, H., Liao, R. (2022). The model of fast face recognition against age interference in deep learning. International Journal of Biometrics, 14(3-4): 223-238. https://doi.org/10.1504/IJBM.2022.124668

[15] Lin, C.H., Huang, W.J., Wu, B.F. (2021). Deep representation alignment network for pose-invariant face recognition. Neurocomputing, 464: 485-496. https://doi.org/10.1016/j.neucom.2021.08.103

[16] Jonas, J., Rossion, B. (2021). Intracerebral electrical stimulation to understand the neural basis of human face identity recognition. European Journal of Neuroscience, 54(1): 4197-4211. https://doi.org/10.1111/ejn.15235

[17] Prabhu, K., SathishKumar, S., Sivachitra, M., Dineshkumar, S., Sathiyabama, P. (2022). Facial expression recognition using enhanced convolution neural network with attention mechanism. Computer Systems Science & Engineering, 41(1): 415-426. https://doi.org/10.32604/csse.2022.019749

[18] Zhang, Y., Liu, W., Fan, H., Zou, Y., Cui, Z., Wang, Q. (2022). Dictionary learning and face recognition based on sample expansion. Applied Intelligence, 52: 3766-3780. https://doi.org/10.1007/s10489-021-02557-2

[19] Zu, F., Zhou, C., Wang, X. (2021). An improved convolutional neural network based on centre loss for facial expression recognition. International Journal of Adaptive and Innovative Systems, 3(1): 58-73. https://doi.org/10.1504/IJAIS.2021.117903

[20] Lei, Y. (2022). Research on microvideo character perception and recognition based on target detection technology. Journal of Computational and Cognitive Engineering, 1(2): 83-87. https://doi.org/10.47852/bonviewJCCE19522514