Journal of Korea Multimedia Society Vol. 24, No. 5, May 2021(pp. 701-708)
https://doi.org/10.9717/kmms.2021.24.5.701

Human Face Recognition Based on improved CNN Model with Multi-layers

Ruyang Zhang[†], Eung-Joo Lee[††]

ABSTRACT

As one of the most widely used technology in the world right now, Face recognition has already received widespread attention by all the researcher and institutes. It has been used in many fields such as safety protection, surveillance system, crime control and even in our ordinary life such as home security and so on. This technology with today's technology has advantages such as high connectivity and real time transformation. But we still need to improve its recognition rate, reaction time and also reduce impact of different environmental status to the whole system. So in this paper we proposed a face recognition system model with improved CNN which combining the characteristics of flat network and residual network, integrated learning, simplify network structure and enhance portability and also improve the recognition accuracy. We also used AR and ORL database to do the experiment and result shows higher recognition rate, efficiency and robustness for different image conditions.

Key words: Face Recognition, CNN Model, ORL Database, AR Face Database, Residual Network

1. INTRODUCTION

Because of the huge development of face recognition technology field, people from all over the world are leading a much more safety and peaceful life nowadays. As feeling the convenience and joy that this technology bring to us in daily life, the research about how to improve it has been the most-watched topic in science technology field. Because of the recognition system has disadvantages such as long reaction time, high hardware demand and can easily be influenced by different environmental conditions, even if the same person's face looks different under different lighting, posture, facial expressions and other factors. Therefore, in order to reduce the impact of these factors on the efficiency of face recognition, it is necessary to minimize the intra-class variation in the same individual and improve the inter-class variation between different individuals when trying to design the whole system. So that's what we try to change and ameliorate in this paper with new system model. The research of face recognition algorithm based on convolutional neural network has made great progress in the past ten years. After a large amount of data training, the studied model can accurately learn face recognition features, so as to achieve better results in face recognition results[1]. We proposed a multi-layer convolutional neural network model for face recognition processing based on residual unit. We used the average pooling layer instead of the fully connected layer, which makes the network structure simple and highly portable. On the basis of improving the CNN network, a voting method-based integrated learning strategy is used to combine the results of all individual learners into the last result to achieve the face recognition function with more accurate.

※ Corresponding Author : Eung-Joo Lee, Address: 428, Sinseon-ro, Nam-gu, Busan, Korea, TEL : +82-51-629-1143, E-mail : ejlee@tu.ac.kr

Receipt date : May 20, 2021, Approval date : May 26, 2021

[†] Dept. of Information & Communication Engineering, Tongmyong University
(E-mail : dlzry@naver.com)
[††] Dept. of Information & Communication Engineering, Tongmyong University※

2. IMPROVED MULTI-LAYER CONVOLUTIONAL NEURAL NETWORK MODEL

The multi-layer convolutional neural network model proposed in this paper consists of 3 convolutional layer, 3 maximum pooling layer, 1 residual unit layer, 1 average pooling layer and 1 SoftMax classification layer. In the network, the convolutional and pooling layers are consisted by multiple feature maps. Each feature map is composed of multiple neurons and the feature maps of each layer are as input to the next layer. The feature map of the convolutional layer may be related to several feature maps of the previous layer. The specific composition is shown in Fig. 1.

2.1 Convolutional Layer

The convolutional layer simulates the process of extracting some primary visual features from simple cells with local receptive fields through local connection and weight sharing[2]. The above set of the same connection strength is a feature extractor, which appears as a convolution kernel during the operation process, and the convolution kernel value is randomly initialized first, and finally determined by network training.

The improved convolutional neural network we proposed uses two convolution kernels of different sizes to process the face image. The size of the convolution kernel of the first two convolutional layers is 5 × 5, and the step size is 1; the size of the convolution kernel that the last convolutional layer having is 3 × 3, and the step size is 1 as well.

The input of each neuron in the convolutional layer comes from the neuron from a fixed area in the feature map of the previous layer, and the size of the area is determined by the size of the convolution kernel[3]. The m(amount) feature maps of the first convolutional layer are convolved by an input image with m learnable convolution kernels, plus bias, and then obtained by the activation function. The n(amount) feature maps of the next convolutional layer are convolved by the m feature maps of the first pooling layer with n×m convolution kernels, and each m convolution results are combined, plus bias, and then obtained by the activation function. The fundamental operating of convolution layer is shown in Fig. 2.

The formula for this layer is:

$$y_m^{(k)} = f\Big(\sum_{i \in M_{m(k-1)}} \sum_{(p,q) \in K^k} W_{mi(p,q)}^{(k)} \times x_i^{(k-1)}(c+p, r+q) + b_m^k \Big) \tag{1}$$

Among them, $K^k = \{(p,q) \in N^2 \, 0 < p < K_x, 0 < q < K_y\}$; K_x and K_y are the width and height of the convolution kernel ; k represents the current number of layers ; b_m^k is the bias of the neuron on the j-th feature map of the convolutional layer ; $M_{m(k-1)(p,q)}$ is the set of feature maps of the previous layer that establishes a relationship with the j-th feature map of the convolutional layer.

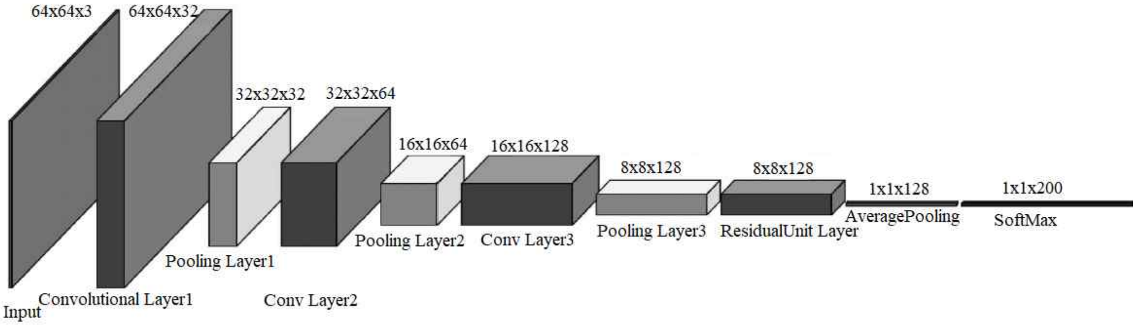

Fig. 1. The Model of the improved Multi-Layer Neural Network.

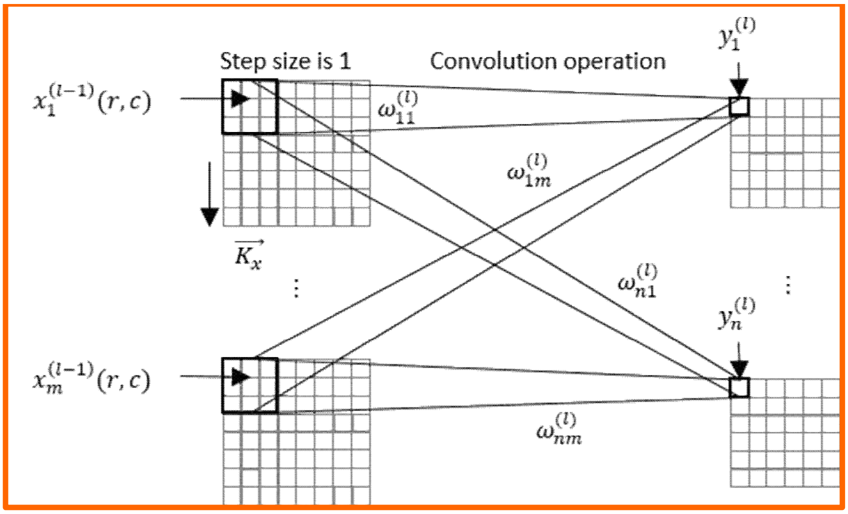

Fig. 2. Convolution layer operation.

2.2 Maximum Pooling Layer

The pooling layer simulates the process by which complex cells screen and combine primary visual features into more advanced, abstract visual features, which are implemented by sampling in the network[4].

Each layer of convolution kernel we designed in this network is followed by a maximum pooling layer. The pooling layer proposed in this paper uses the maximum sampling, and the sampling size is 2×2, which is, dividing the input feature map into non-overlapping 2×2 rectangles, and taking the maximum value for each rectangle, so the length and width of the output feature map are half of the input feature map. The mathematical formula of this layer is:

$$y_m^k = down(y_m^{(k-1)}) \quad (2)$$

Among them, *down* represents a function of maximum sampling, and this layer operation does not include learnable weights and thresholds.

2.3 The Residual Unit

The network algorithm we proposed in this paper combines the normal convolutional neural network and the residual neural network together to make the improved neural network, in which the structure is simpler, also there are less parameters, so it has a faster convergence speed at the same time.

This unit is a 3-layer network structure, including 2 convolutional layers with convolution kernel which is in size of 1 × 1 and 1 convolutional layer with convolution kernel which is in size of 3 × 3. The residual unit directly maps the low-level image features to the high-level network through short connections. This method detours the input information and outputs it directly in order to ensure the integrity of the information. The structure of the residual unit is shown in Fig. 3.

The residual unit first uses a 1 × 1 convolution kernel to reduce the dimensionality of the input feature vector, so that the output feature vector becomes 64-dimensional, and then uses a 3 × 3 convolution kernel to convolve the feature vector. In order to align the dimensions of the input vector and the output vector, the feature vector we got needs to go through a 1 × 1 convolution kernel again to increase the dimension. Add the input vector and output vector of the residual unit and do processing of a non-linear operation, and then use it as the input of the average pooling layer.

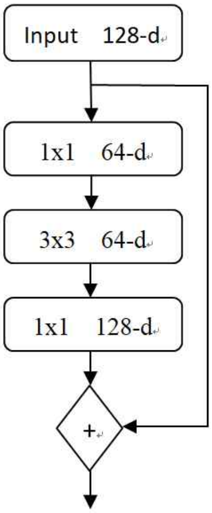

Fig. 3. The Structure of residual unit.

The average pooling layer reduces features and parameters while maintaining rotation, translation, expansion, etc. And then we can get the data we need to use as an input to SoftMax layer.

2.4 Softmax Regression Layer

Because the face features are more complex than digital features, and there are no uniform templates for face types[6]. So at the last layer of the network we uses Softmax regression with strong nonlinear classification ability as the classifier.

Assume that m samples that can be divided into k categories compose the training set $\{(x^{(l)},y^{(l)}),\dots,(x^{(m)},y^{(m)})\}$, sample $x^{(i)}\in R^{(n+1)}$, class mark $y^{(i)}\in\{1,2,3,\dots,k\}$, n is sample dimension so the Softmax regression hypothesis function is:

$$h_\theta(x^{(i)}) = \begin{bmatrix} p(y^{(i)}=1|x^{(i)};\theta) \\ \vdots \\ p(y^{(i)}=k|x^{(i)};\theta) \end{bmatrix} = \begin{bmatrix} \exp(\theta_1^T x^{(i)}) \\ \vdots \\ \exp(\theta_k^T x^{(i)}) \end{bmatrix} / \sum_{j=1}^{k} \exp(\theta_j^T x^{(i)}) \tag{3}$$

Among them, $p(y^{(i)}=k|x^{(i)};\theta)$ represents the probability that the sample $x^{(i)}$ belongs to the class j; $\theta_j^T\in R^{(n+1)}$ represents model parameters. Written $\theta_1,\theta_2,\cdots,\theta_j$ in matrix form by row $\theta\in R^{k(n+1)}$. The cost function of the model is:

$$J(\theta)=\frac{-1}{m}\left[\sum_{i=1}^{m}\sum_{j=1}^{k} l\{y^{(i)}=j\}\log\frac{e^{\theta_j^T x^{(i)}}}{\sum_{l=1}^{k} e^{\theta_l^T x^{(i)}}}\right] \tag{4}$$

Among them, $l\{\bullet\}$is indicative function; If the internal value of {} is a true expression, the indicative function value is 1; otherwise, it is 0. In this paper, the batch gradient descent method is used to solve the Θ when the cost function $J(\theta)$ is minimized.

3. Improved CNN network combined with integrated learning

The ensemble learning algorithm is one of the most popular machine learning algorithms nowadays. It completes the learning task by constructing and combining several individual learners. The ensemble learning combines multiple learners to generally obtain better generalization performance than a single learner. There are two commonly used combining strategies which are voting method and learning method. In this paper we used the voting method.

We proposed a face recognition algorithm based on improved CNN network and ensemble learning, which is a homogeneous ensemble algorithm. The algorithm is composed of two individual learners which are two simple convolutional neural networks. The first convolutional neural network in the improved CNN network consists of 3 convolutional layers, 3 pooling layers, 1 residual unit layer and 1 fully connected layer and 1 Softmax layer; The second convolutional neural network contains 3 convolutional layers, 3 pooling layers, 1 residual unit layer, 1 average pooling layer, and 1 SoftMax layer. The feature extraction layers of the two networks are different, one is a fully connected layer, and the other is an average pooling layer, which makes the face feature vectors that we got as output have differences. The two networks are both independent at the same time interrelated. Both use the cross-entropy function as the network loss function, and form a new loss function based on their own loss functions.

4. EXPERIMENT RESULT

4.1 Data preprocessing

"min-max" normalize the gray value, that is, normalize the gray value of each pixel in the picture to [0, 1]. Use x and x' to represent the current grayscale values and normalized grayscale values, min and max represent the minimum and maximum grayscale values in the picture[7]. The standardized formula is :

$$x' = \frac{x - \min}{\max - \min} \tag{5}$$

4.2 Network training algorithm

Due to the large number of face database training samples, this paper uses a batch random gradient descent method which has faster convertgence in practice[8]. For the training of the convolutional neural network of the ORL face database, the batch block size is 40, the momentum is 0.9, and the learning rate is 0.12 constant. For the training of the convolutional neural network of the AR face database, the batch block size is 65, the momentum is 0.9, and the learning rate is 0.15 constant. Each iteration traverses all the batch blocks of the training set, and updating the network parameters once as soon as traverses a batch block each time[9]. The update formula is:

$$w_{i+1} = \epsilon \cdot w_i - \eta \cdot \left(\frac{\partial L}{\partial w_i} \right)_{D_i} \tag{6}$$

Among them, w_i is the current parameter, w_{i+1} is the updated parameter, ϵ is momentum, η is Learning rate. $\left(\frac{\partial L}{\partial w_i} \right)_{D_i}$ is the average value of the partial derivative of error to w_i in the i-th batch processing block D_i.

4.3 Database Introduction

The experimental data comes from the ORL, AR face databases. There are 600 different lighting conditions in the ORL database, different poses,

Fig. 4. Part of the image in ORL Database.

Fig. 5. Part of the image in AR Database.

and different facial expressions. Each person has 12 images totaling 50 people. Compared to ORL face database, AR face database is a large-scale color face image database. The database has a total of 4,000 images with 768 × 576 pixels, including 126 people. Each person in the library has 8 images under normal lighting conditions, 6 images under varying light conditions, and 4 images with different expression changes[12]. The samples of those two face databases are as follows, first is the ORL face database:

Then there is the AR face database:

4.4 Experiment Method

While experimenting, the size of the convolution kernel in the first 2 convolution layers is 5×5, the third one is 3×3, and the 3 pooling layers utilize maximum pooling, the sampling size is 2×2, and the activation function utilizes the sigmoid function, which is:

$$f(x)=\frac{1}{1+e^{-x}} \tag{7}$$

The inputs of the network model are 64 × 64 × 3's RGB color human face images. During the experiment processing, we used the set-out method to keep the experiment running well, and find the average value through multiple experiments at the same time. In order to make the experimental results more accurate, the data set was randomly shuffled in each experiment, and then 80% of the data was used as the training set and the rest 20% as the test set.

The network weight is initialized by Gaussian, and the network threshold is initialized to 0, the convolutional layer, the average pooling layer and the Softmax regression layer all use the "dropout" method, and the probabilities p are set to 0.5, 0.2 and 0.5 respectively.

Compared the proposed CNN model with other algorithm based on both ORL and AR face image database, the results are shown in the Table 1 and Table 2 below.

Table 1. Proposed model compared with other algorithms based on ORL Database.

Recognition Methods	Recognition Rate (%)
Eigenface	97.15
ICA	94.36
2DPCA	98.47
Improved CNN Model	99.74

Table 2. Proposed model compared with other algorithms based on AR Database.

Recognition Methods	Recognition Rate (%)
PCA	85.53
2DPCA	94.98
PCA+GSRC	97.66
Improved CNN Model	99.43

Through the results we can see that the network system proposed in this paper has a better performance based on both face database and the network has stronger ability to resist the interference of illumination, facial expression changes, and whether there is any obstruction.

5. CONCLUSION

For small and medium-sized face databases, this paper proposed a Multi-layer Convolutional Neural Network System Model combined with the residual network and integrated learning and used it to do experiment with both ORL and AR face database. The improved CNN Network model has the advantages that both CNN and residual network have, which are fast convergence speed and simple structure. And combined with integrated learning, we can improve the recognition rate to a further step. The recognition rate for all the testing samples in those two databases are 99.74% and 99.43%. When it was used in AR face database the result shows its robustness to impact such as illumination differences, facial emotional changes, and the presence or absence of obstructions. And also when we did the experiment based on the environment of Microsoft system and Matlab, the face recognition time is pretty fast and the system is very stable, so the real-time recognition consequent is good. In the future, we will keep improving the network structure, and combining different functions to improve the generalization performance of the network model and the accuracy of the whole system to make it fitting more different situation and environment.

REFERENCE

[1] R.Y. Zhang and E.J. Lee, "Face Recognition Based on Improved Convolutional Neural Network," *Proceeding of the Spring Conference of the Korea Multimedia Society*, pp. 8-10, 2019.

[2] R Chellappa, C L Wilson, and S Sirohey, "Human and machine recognition offaces: A survey," *Proceedings of the IEEE*, Vol. 83, No. 5, pp. 705-740, 1995.

[3] H.H. Nam, B.J. Kang, and K.H. Park. "Comparison of Computer and Human Face Recognition According to Facial Components," *Journal of Korea Multimedia Society*, Vol. 37, No. 21, pp. 40-50, 2012.

[4] A.R. Syafeeza, M. Halil-Hani, S.S. Liew, and R. Bakhteri, "Convolutional Neural Network for Face Recognition with Pose and Illumination Variation," *International Journal of Engineering & Technology*, Vol. 6, No. 1, pp. 44-57, 2014.

[5] A. Toshev and C. Szegedy, "Deeppose: Human Pose Estimation via Deep Neural Networks," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos*, pp. 1653-1660, 2014.

[6] S. Srivastava, G. Hintion, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Over-fitting," *Journal of Machine Learning Research*, Vol. 15, No. 1, pp. 1929-1958, 2014.

[7] A.F. Abate, M. Nappi, D. Riccio, et al., "2D and 3D Face Recognition a Survey," *Journal of Pattern Recognition Letters*. Vol. 28, No. 14, pp. 1885-1906, 2007.

[8] P. Sermanet, D. Eigen, and X. Zhang. "Overfeat Integrated Recognition, Localization and Detection Using Convolutional Network," *Journal of Neural Networks*, Vol. 16, No. 5, pp. 555-559, 2003.

[9] F. Hajati, M. Tavakolian, S. Gheisari, G. Yongsheng, and A.S. Mian, "Dynamic Texture Comparison Using Derivative Sparse Representation: Application to Video-based Face Recognition," *Journal of IEEE Transactions on Human-Machine Systems*. Vol. 47, No. 6, pp. 970-982, 2017.

[10] M.Y. Liu, S.G. Shan, R.P. Wang, and X.L. Chen, "Learning Expressionlets via Universal Manifold Model for Dynamic Facial Expression Recognition," *Journal of IEEE Transactions on Image Processing*, Vol. 25, No. 12, pp. 5920-5932, 2016.

[11] T.M. Guo, J.W. Dong, H.J. Li, and Y.X. Gao, "Simple Convolutional Neural Network on Image Classification," *Proceedings of IEEE 2nd International Conference on Big Data Analysis(ICBDA)*, pp. 721-724, 2017.

[12] R.Y. Zhang and E.J. Lee, "Face Recognition Based on 6-layer CNN Model," *Proceeding of the Spring Conference of the Korea Multimedia Society*, pp. 333-335, 2020.

Ruyang Zhang

received his B.S. at Dalian Polytechnic University in China (2012-2016) and Master Degree in Tongmyong University(2016-2018). Currently he is studying in the Department of Information and Communication Engineering in Tongmyong University, Korea for his Doctor degree. His main research areas are image processing and face recognition.

Eung-Joo Lee

received his B. S., M. S. and Ph. D. in Electronic Engineering from Kyungpook National University, Korea, in 1990, 1992, and Aug. 1996, respectively. Since 1997 he has worked with the Department of Information & Communications Engineering, Tongmyong University, Korea, where he is currently a professor. From 2005 to July 2006, he was a visiting professor in the Department of Computer and Information Engineering, Dalian Polytechnic University, and from Dec 2018 he was appointed honorary professor of Dalian Polytechnic University, China. His main research interests include biometrics, image processing, and computer vision.