Journal of Korea Multimedia Society Vol. 21, No. 12, December 2018(pp. 1467-1472)
https://doi.org/10.9717/kmms.2018.21.12.1467

Low Resolution Rate Face Recognition Based on Multi-scale CNN

Ji-Yuan Wang[†], Eung-Joo Lee[††]

ABSTRACT

For the problem that the face image of surveillance video cannot be accurately identified due to the low resolution, this paper proposes a low resolution face recognition solution based on convolutional neural network model. Convolutional Neural Networks (CNN) model for multi-scale input The CNN model for multi-scale input is an improvement over the existing "two-step method" in which low-resolution images are up-sampled using a simple bi-cubic interpolation method. Then, the up sampled image and the high-resolution image are mixed as a model training sample. The CNN model learns the common feature space of the high- and low-resolution images, and then measures the feature similarity through the cosine distance. Finally, the recognition result is given. The experiments on the CMU PIE and Extended Yale B datasets show that the accuracy of the model is better than other comparison methods. Compared with the CMDA_BGE algorithm with the highest recognition rate, the accuracy rate is 2.5%~9.9%.

Key words: Face Recognition; Intelligent Video Analysis Method; CNN; Multi-scale CNN

1. INTRODUCTION

Vision is the main way that humans perceive the world. The information obtained by the human brain through the visual system accounts for more than 80% of all sensory information. To realize the intelligence of the video surveillance system, it is necessary to let the surveillance system not only passively provide video data, but also can automatically analyze and process the video content. The computer vision technology is to realize the surveillance system from “visible” to “seeing clearly” The key technology for change [1]. By continuously improving the video analysis technology, the monitoring system can accurately and timely detect abnormal behavior in the monitoring picture in a large amount of video data, give the alarm in the fastest and best way, and identify the suspicious target. Realize automatic intelligent monitoring.

2. RELATED WORK

2.1 Intelligent video analysis method

Video Analysis (VA), also known as Video Content Analysis (VCA), the industry will use video analytics technology video surveillance called intelligent video surveillance (Intelligent Video Surveillance IVS).

Intelligent video analysis technology was born in the 90s of last century. Once this technique originated in computer vision, it has attracted the attention of some advanced scientific research institutes abroad and set off an upsurge of research. The VSAM system[2] belongs to the early developed intelligent monitoring system, which can only identify and track simple targets such as individuals and vehicles. Afterwards, the ADVISOR

※ Corresponding Author : Eung-Joo Lee, Address: 428, Sinseon-ro, Nam-gu, Busan, Korea, TEL : +82-51-629-1143, FAX : +82-, E-mail : yuan212108@naver.com, ejlee@tu.ac.kr
Receipt date : Oct. 8, 2018, Approval date : Oct. 31, 2018

[†] Dept. of Information and Communication Engineering, Tongmyong University
(E-mail : yuan212108@nver.com,ejlee@tu.ac.kr)
[††] Dept. of Information and Communication Engineering, Tongmyong University

system[3], funded by the EU Information Society Technology Project, already has the functions of target tracking, crowd analysis and behavior analysis, and enables intelligent real-time monitoring of subway stations. The S3 system developed by IBM[4] is an open platform architecture. Different video analysis modules can be integrated into the system to provide users with a full range of monitoring and data analysis services.

2.2 Deep learning method

Deep learning is a multi-layer feature learning method[5-7]. It learns the law from massive data in an end-to-end manner with the aid of a deep neural network. Deep learning has achieved remarkable results in the areas of image processing, speech recognition and natural language processing, which has greatly accelerated the development of machine learning and has become a new milestone in the history of artificial intelligence. In the process of realizing intelligent video analysis, a series of image processing and pattern recognition problems need to be dealt with, and deep learning methods provide a powerful theoretical basis for solving these problems[13].

2.3 Convolutional Neural Network

Convolutional neural network (CNN) is another discriminative model widely used in image processing and speech recognition[8]. Because of the high dimensionality of the image data, if a neural network is constructed in a fully connected manner, it will certainly bring about an abnormally large number of parameters. A convolutional neural network is a deep neural network specially designed for processing two-dimensional data.

In order to reduce the complexity of the model, the CNN model adopts three special processing methods when constructing. 1) Local Receptive Field, 2) Weight Sharing, 3) Pooling operation. Through the above three methods, the CNN model reduces the parameters while making the extracted image features have the characteristics of displacement invariance, scale invariance, and rotation invariance. The training process of the CNN model is similar to the traditional back propagation algorithm. It is mainly divided into the forward propagation phase and the backward propagation phase. The training of the CNN model is a supervised training. Therefore, a large amount of data samples are needed to support the massive samples the CNN model can often achieve excellent performance.

3. DEEP NETWORK STRUCTURE BASED ON MULTI-SCALE INPUT CNN

The low-resolution face recognition problem can be expressed as follows: For a low-resolution face image and an image search library containing N high-resolution images , our goal is to use the sample library. The identity of is identified in . Due to the high dimensionality of the image data, it is often necessary to perform face matching through the extracted features. Assuming and are two feature extraction maps, the objective function can be expressed as:

$$\min_i dist(P(I_L), Q(I_{Hi})) \quad i=1,2,...,N \qquad (1)$$

Where $dist(\bullet)$ is the distance function used to measure the similarity between features $P(I_L)$ and $Q(I_{Hi})$. When P and Q are the same matrix and defined as F , because of the input dimensions of the high-resolution and low-resolution images are different, it is necessary to upsample the low-resolution images to obtain a generated high-resolution image I_H. So the objective function is rewritten as:

$$\min_i dist(F(I_H), F(I_{Hi})) \quad i=1,2,...,N \qquad (2)$$

This paper uses deep learning methods to solve the problem of face recognition in low resolution. Based on the "two-step method", this paper proposes a multi-resolution input network architecture based on multi-resolution convolutional neural networks (CNN). The biggest difference in the two-step method is that the MSCNN model proposed in this paper is more concerned with the step of feature expression.

The entire system can be divided into two phases, as shown in Fig. 1. During the training phase, bi-cubic interpolation upsampling of the low resolution is first required to ensure that the input image has the same size. After preprocessing, different resolution images are mixed to obtain a training sample set. Through the classification and training of these different images, a convolutional neural network model can be obtained, and then the model can be used to extract features from the sample library image and obtain a feature sample library. In the test phase, the low-resolution image is first upsampled by bi-cubic interpolation, and then the trained model is used for feature extraction. Then the nearest neighbor classifier is used to measure the feature similar to the feature library. Degree, the tag corresponding to the feature with the smallest cosine distance is selected as the identification category of the test image.

This paper selects the Convolutional Neural Network (CNN) in the depth model for feature extraction. Using a pre-processed tagged multi-resolution image, a multi-scale input convolutional neural network model (MSCNN) can be trained. By training the MSCNN model, it is possible to achieve clustering of different resolution images of the same person, thus indirectly ensuring the similarity of intra-class image features and the differences of image features between classes.

The structure of the MSCNN model is shown in Fig. 2. The input to the network is a mixed gray image with different resolutions. The input size is 1×32×32. The two convolutional layers are connected to two maximum pooling layers to extract the local information of the image. The convolution kernel size is 5×5, the downsampling factor is set to 2, and the ReLU function is used as a nonlinear activation function. After convolution, 64 maps of 5×5 feature can be obtained. These feature maps

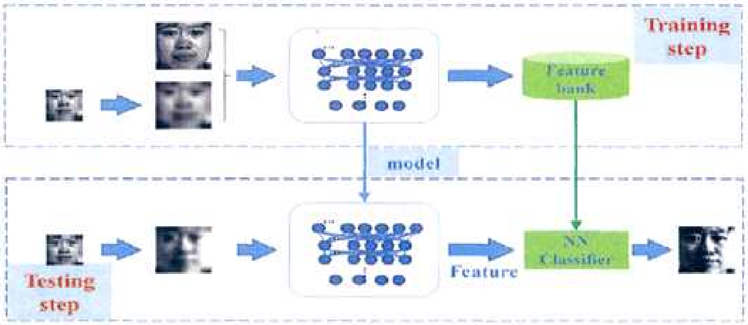


Fig. 1. Multi-scale Convolutional Neural Network (MSCNN) algorithm training and testing process.

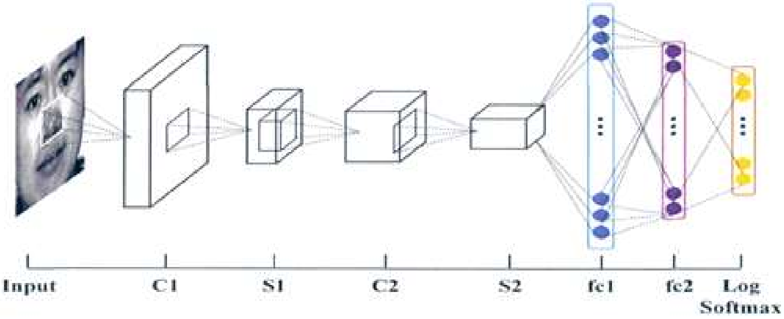


Fig. 2. Multi-scale convolutional neural network structure.

are vectorized and input into the fully connected layer fc1 for further feature fusion. These features are also used as the final classification. Finally, the features produced by fc1 are categorized by fully connecting the layers fc2 and LogSoftmax layers.

The parameter θ of the model can be optimized by minimizing the difference between the model prediction and the corresponding tag value. In this paper, the negative log-likelihood function is chosen as the cost function of the model. For a given image set X_i and corresponding label information Y_i, the negative likelihood logarithm is defined as:

$$L(\theta) = \sum_{i}^{n} \sum_{k}^{K} - Y_i^k F(X_i; \theta) \tag{3}$$

Where K is the total number of categories and n represents the number of samples in each category.

In this paper, the back-propagation algorithm is used, and the stochastic gradient descent is used to optimize the parameters and update so that the cost function is minimized. When training, select Batchsize as 10, that is, perform iterative processing for every 10 samples. Select 0.99 for momentum and 0.00001 for weight attenuation. The weight w update strategy is shown in Equation 4:

$$\begin{aligned} v_{i+1} &:= 0.99 \bullet v_i - 0.00001 \bullet \epsilon w_i - \frac{\partial L}{\partial w}\Big|_{wi} \\ w_{i+1} &:= w_i + v_{i+1} \end{aligned} \tag{4}$$

Where i is the number of iterations index, v is the momentum parameter, and ϵ is the learning rate, set to 0.0010 at the start of the experiment.

In order to prevent the model from being over-fitted due to lack of training samples in the training process, two regularization methods are used in this paper data expansion and Dropall method, which uses the DropOut[9] and DropConnect[10] methods together and samples randomly by discarding some nodes or neuron connections. Through this method, the interdependence between neurons is reduced, so that the model can learn more robust features to reduce over-fitting.

4. EXPERIMENTAL CLASSIFICATION RESULTS AND ANALYSIS

In order to verify the accuracy of the model, experiments were performed on the public data sets CMU PIE[11] and Extended Yale B[12]. In this paper, each volunteer was selected to experiment with five images with different facial expressions and lighting under five close postures (C05, C07, C09, C27, and C29). The Extended Yale B contain 9 different poses and 64 different lighting changes. This article also selected 2,432 frontal images under all lighting conditions for experiments.

In all experiments, the high-resolution samples were generated from the face images of the dataset and were trimmed and cropped to a size of 32 × 32 pixels. By downsampling these high-resolution images, corresponding low-resolution images can be obtained, with a low-resolution size of 12 × 12 pixels. In this paper, three sets of different experiments were performed on each data set. In the experiment, 10, 20, and 30 pairs of high and low resolution image pairs were selected to form a sample set to train the model, and the remaining low-resolution images were used. Test model accuracy. The model parameter settings in the experiment are shown in Table 1.

First, this section tests the recognition perform-

Table 1. Layer parameter list

Model	Convolution kernel	Output size
Input		1*32*32
C1	5*5/0/1	32*28*28
DropOut(50%)		32*28*28
S2	2*2/0/2	32*14*14
C2	5*5/0/1	64*10*10
DropOut(50%)		64*10*10
S2	2*2/0/2	64*5*5
fc1		1*1*150
DropOut(50%)		1*1*150
DropConnect(50%)		1*1*150
fc2		1*1*16(38)
LogSoftmax		1*1*68(38)

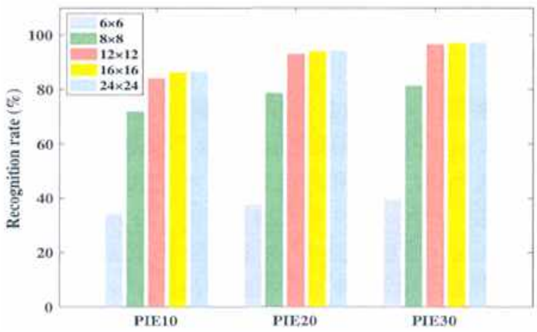


Fig. 3. Different resolution face recognition results on CMU PIE database.

ance of the algorithm on the CMU PIE data set. The result is shown in Fig. 3. The MSCNN model achieved 84.04%, 92.95%, and 96.02% recognition accuracy in the training samples for 10, 20, and 30, respectively. It is worth noting that with the increase in the number of samples, the accuracy of the MSCNN model has been greatly improved, which also shows that the MSCNN model can achieve better results in the case of large amounts of data.

5. CONCLUSION

This paper examines the problem of face recognition at low resolution. To solve this problem, this paper proposes a face recognition models based on convolutional neural networks with the help of deep learning and powerful feature expression capabilities a multi-scale input-based CNN model using simple bicubic interpolation for low-resolution images. Experiments show that the face recognition model based on convolutional neural network can express the characteristics of face image under low resolution well, so as to achieve a very advanced recognition effect.

REFERENCE

[1] W.K. Xu and E.J. Lee, "Human-computer Catural User Interface Based on Hand Motion Detection and Tracking," *The Journal of Multimedia Information System*, Vol. 15, No. 4, pp. 501-507, 2012.

[2] R.T. Collins, A.J. Lipton, and T. Kanade, A System for Video Surveillance and Monitoring, *Vsam Final Report Carnegie Mellon University Technical Report*, 2000.

[3] Siebel, T. Nils, and S.J. Maybank, "The Advisor Visual Surveillance System," *Proceeding of European Conference on Computer Vision 2004 Workshop Applications of Computer Vision*, pp. 103-111, 2004.

[4] C.F. Shu, A. Hampapur, M. Lu, L. Brown, J. Cannell, A. Senior, Y.L. Tian, et al., "IBM Smart Surveillance System (S3): a Open and Extensible Framework for Event based Surveillance," *Proceeding of IEEE Conference on Advanced Video and Signal Based Surveillance*, pp. 318-323, 2005.

[5] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *Computer Science*, 2014.

[6] E. Ahmed, M. Jones, and T.K. Marks, "An Improved Deep Learning Architecture for Person Re-identification," *Computer Vision and Pattern Recognition*, pp. 3908-3916, 2015.

[7] K.M. He, X.Y. Zhang, S.Q. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *arXiv*, arXiv:1512.03385, 2015.

[8] K.T. Lim, H.W. Kang and J.K. lee,"Moving Shadow Detection using Deep Learning and Markov Random Field," *journal of multimedia information system*, Vol. 18, No. 12, pp. 1432-1438, 2015

[9] G.E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R.R. Salakhutdinov, "Improving Neural Networks by Preventing Co-adaptation of Feature Detectors," *Computer Science*, Vol. 3, No. 4, pp. 212-223, 2012.

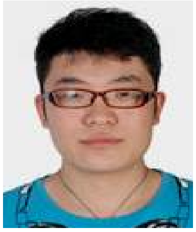
[10] L. Wan, M.D. Zeiler, S.X. Zhang, Y. Lecun, and R. Fergus, "Regularization of Neural Networks using DropConnect," *Proceeding of International Conference on Machine Learn-*

ing, pp. 1058-1066, 2013.

[11] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-PIE," *Image and Vision Computing,* Vol. 28, No. 5, pp. 807-813, 2010.

[12] A.S. Georghiades, P.N. Belhumeur, and D.J. Kriegman, "From Few to Many: Illumination Cone Models for Face Recognition under Variable Lighting and Pose," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. 12, No. 6, pp. 643-660, 2002.

[13] F. Rosenblatt, "The Perception: a Probabilistic Model for Information Storage and Organization in the Brain," *American Psychological Association*, Vol. 65, No. 6, pp. 386-408, 1958.



Jiyuan Wang

received his B.S. at Tongmyong University in Korea (2010-2015) and Master degree at Tongmyong University in Korea (2015-2017). Currently, he is studying in Department of Information and Communications Engineering Tongmyong University for PH.D. His main research areas are image processing, computer vision, biometrics and pattern recognition.

Eung-Joo Lee

received his B. S., M. S. and Ph. D. in Electronic Engineering form Kyungpook National University, Korea, in 1990, 1992, and Aug. 1996, respectively. Since 1997 he has been with the Department of Information & Communications Engineering, Tongmyong University, Korea, where he is currently a professor. From 2000 to July 2002, he was a president of Digital Net Bank Inc. From 2005 to July 2006, he was a visiting professor in the Department of Computer and Information Engineering, Dalian Polytechnic University, China. His main research interests include biometrics, image processing, and computer vision.