Face Detection and Recognition from Distance Based on Deep Learning

Hui Wang and Wei Qi Yan
Auckland University of Technology, Auckland 1010 New Zealand

ABSTRACT

Face recognition is an important biometric in video surveillance. However, the conventional algorithms of face recognition are susceptible to various conditions. The contribution of this chapter is to explore human face recognition by using deep learning methods, including positioning human faces on the given images at various distances and multiple angles. In the distances, the influence from the camera to a face and the size of the face in the images is diverse. We have collected various videos as the input and applied them to train the proposed models. The accuracy of human face recognition from the videos excluding training dataset is at 90.18%. The results indicate that deep learning method is able to recognize human faces with partial occlusion and various distances.

Keywords: Face Detection, Face Recognition, Deep Learning, Distance, Partial Occlusion, Face Size, Accuracy

INTRODUCTION

With the rapid development of computer vision, artificial intelligence (AI) has become the core of contemporary high-tech, its applications include face detection and recognition. The key issue in face detection and recognition is feature extraction which has been developed rapidly based on computational intelligence (Bonetto *et al*. 2015; Wang & Srinivasan, 2017). There are many types of applications of face recognition, e.g., gender identification, ages and emotions are the most important characteristics of human faces. The main purpose of this book chapter is to automatically recognize human faces and affirm the class of the objects. This topic is also one of the important research fields in computer vision and deep learning (LeCun, Bengio, & Hinton, 2015).

Face recognition, as a biometric, is a significant components of video surveillance and visual security which has been applied to human identification nowadays. In large shopping malls, face recognition is employed to monitor the passengers and provide users with convenient services. At the entrance of a railway station, airport, bank, supermarket, school or company, face recognition is applied to implement access control, which prevents the entry of aliens and ensures the security of premise (Zhang, *et al*. 2017). With the development of human face recognition, it has been applied to protect the privacy of users, improve the security of visual data. Face recognition, as a special part of human-computer interaction (Bian, *et al*. 2016) identifies users, serves them with great convenience.

In 1943, McCulloch and Walter (Landahl, McCulloch, & Pitts, 1943) proposed that the first artificial neuron model: MP model who connected the basic unit together to understand how human brain produces highly complex patterns. This has made a significant contribution to the development of artificial neural networks. In 1958, Rosenblatt greatly developed the neural network theory and applied it to real problems. In 1986, Rumelhart et al. proposed backpropagation algorithm, which is an important method in the neural network to calculate the errors of neurons after data processing (Liu, & Liang, 2005). This algorithm is still the most popular one of the most broadly applied artificial neural networks in artificial intelligence.

With the development of neuroscience, computer science scientists have found that brain signals are transmitted through a complex structure; if time permits, the characteristics are applied to understand digital signals, which led to the emergence of deep learning (LeCun, Bengio, & Hinton, 2015) for the establishment and simulation of human brain for analysis and learning. Convolutional neural networks (CNNs) (Karpathy, *et al*. 2014) have been successfully applied to visual imagery in the past few years. One of the most important factors is the need to provide a large amount of training data. But in face recognition, due to lack of large scale of data sets, a few of experiments were limited.

The contribution of this book chapter is based on deep learning for face recognition, which will be completed in real time. For example, if people are near to the camera, the system will verify the influence of proportion of face to the total size of the given images, the proportion is thought as a major core. The relevant experiments require four parts: 1) Collecting the data set, 2) accepting the command parameters, and 3) defining the neural network model, 4) training and testing model.

This book chapter is organized as follow. Literature review is presented after this introduction section, then the method and final results are demonstrated, the conclusion is drawn at last.

RALATED WORK

Deep Neural Networks (DNNs) have a history of more than 10 years. The DNNs were originated in 1943 when Pitts and McCulloch (McCulloch, & Pitts, 1943) created a computer model based on human brain neural network that eventually became a hot topic after a century of development. Neocognitron (Fukushima, & Miyake, 1982) was the first artificial neural network that introduced CNNs, where the receptive field of a convolutional unit gave weight vector. In 2006, Hinton (Hinton, Osindero, & Teh, 2006) presented the concept of a fast learning algorithm for deep belief net, who presented the deep learning methods and improvement of DNNs training model. However, deep learning remains in theoretical stage till Hinton's team won the championship of ImageNet Large Scale Visual Recognition Challenge (ILSVRC) (Simonyan, 2015) by using AlexNet in 2012, it is an important milestone in the history (Schmidhuber, 2015). The deep learning has aroused widespread concern after the ImageNet award.

There are four reasons (Simonyan, 2015) that why AlexNet succeeded and DNN became one of the most popular topics: 1) Big dataset with millions ImageNet data, 2) assisting with GPU acceleration for model training, 3) the methods of preventing overfitting, e.g., dropout and data augmentation; 4) the development of nonlinear activation function, e.g., ReLU, tan, sigmoid.

The number of images in dataset is an important factor that determines the accuracy of visual object recognition in deep learning. The method of data augmentation is to enhance the training data by using mathematical transformations such as Affine transformation, in order to achieve the purpose of reducing overfitting and improving the accuracy of recognition results. The two distinct forms of data augmentation respectively are to generate image translations and horizontal reflections and to alter the intensities of the RGB channels in training images (Krizhevsky, Sutskever, & Hinton, 2012). Both methods are able to convey multiple images from the original one with very few computations; the transformed images do not need to be stored on disk which are generated before model training in deep learning.

Dropout was presented by Hinton (Srivastava, *et al*, 2014) in 2012. The dropout is a neural network unit temporarily discarded from deep learning models in accordance with a probability in the training process of deep learning. It should be noted that the temporarily discarded neuron parameters are merely hidden in this training phase; the essence is to ignore the part of feature classifier, which means that the part of hidden layer nodes tends to zero in each loop of training. This approach can reduce the interaction in feature classifier, thus, effectively diminish the overfitting phenomenon.

Dropout is broadly employed in the field of DNNs (Gal & Ghahramani, 2016), which directly shows the superiority of this method in improving the accuracy of verification results. But the drawbacks of dropout are also noteworthy, which greatly increase the time of data training and the complexity of the nonlinear activation function (Maalej, Tagougui, & Kherallah, 2016). However, both dropout and data augmentation are effective methods to reduce overfitting.

Multilayer perceptron (MLP) is interpreted as an artificial neural network, which contains input layer, hidden layer, and output layer. The layers are connected by using fully-connected layer in MLP; the simplest MLP can have only one hidden layer. In CNNs, an MLP consisting of multiple fully-connected layers with nonlinear activation functions (Lin, Chen, & Yan, 2013). We see that images input to the MLP layer (Bengio, Courville, & Vincent, 2013) in the CNNs are an abstract one from convolution layers which describes that this feature is invariant to variations of the same concept after the feature extraction from convolutional layers.

In MLP, the function is described (Gal & Ghahramani, 2016), if the input vector is x, the output vector is $f(x)$, bias vectors are $b^{(1)}$ and $b^{(2)}$, weight matrices are $W^{(1)}$ and $W^{(2)}$, activation functions are $G(\cdot)$ and $s(\cdot)$, then the matrix notation of MLP is,

$$f(x) = G(b^{(2)} + W^{(2)}(s(b^{(1)} + W^{(1)}x))). \quad (1)$$

The output vector of MLP is obtained,

$$o(x) = G(b^{(2)} + W^{(2)}h(x)) . \quad (2)$$

In order to train the MLP layer, the DNN model needs to be trained with all the parameters, the set of parameters is $\{W^{(2)}, b^{(2)}, W^{(1)}, b^{(1)}\}$.

MATHEDOLOGY

In this paper, we collected face images as the training dataset for five classes. The images with different distances, environments with various lighting conditions, human faces with multiple angles were taken as the test dataset, which increases the difficulty of face detection and recognition. Figure 1 shows the dataset.

In order to improve the accuracy of face recognition, the amount of original data was augmented by generating new training dataset using the existing images. In this project, an image is generated 50 images after data augmentation like random cropping and scaling from multiple angles, random cropping is to enrich the face size and information, the position and size of the face have been modified after the randomly cropping. Figure 2 shows an image after randomly cropping, scaling, and rotating.

In deep learning, the use of CNNs to achieve face recognition has attracted broad attention. This chapter refers to the SSD (Liu, *et al*. 2016) architecture for model design so as to improve the speed of face recognition. Firstly, the input image is extracted through a CNN architecture. Then, the model takes advantage of two MLP (Cireşan, *et al*., 2011) structures to implement visual object classification and localization respectively as the input of CNN model. Fig. 3 is the diagram of model structure.

At first, if the DNN model is trained by using the input images, it is not an image followed by another, but N images will be input in one step (Krizhevsky, Sutskever, & Hinton, 2012), N is set as batch size. At the beginning of this model, the input images are needed, the dimension of model parameters is 4D. Among

them, 360×640 is the size of the image, 3 is the number of colour channels of the image, because the image is with RGB 3-color channels (Wu, Lin & Chang, 2007).

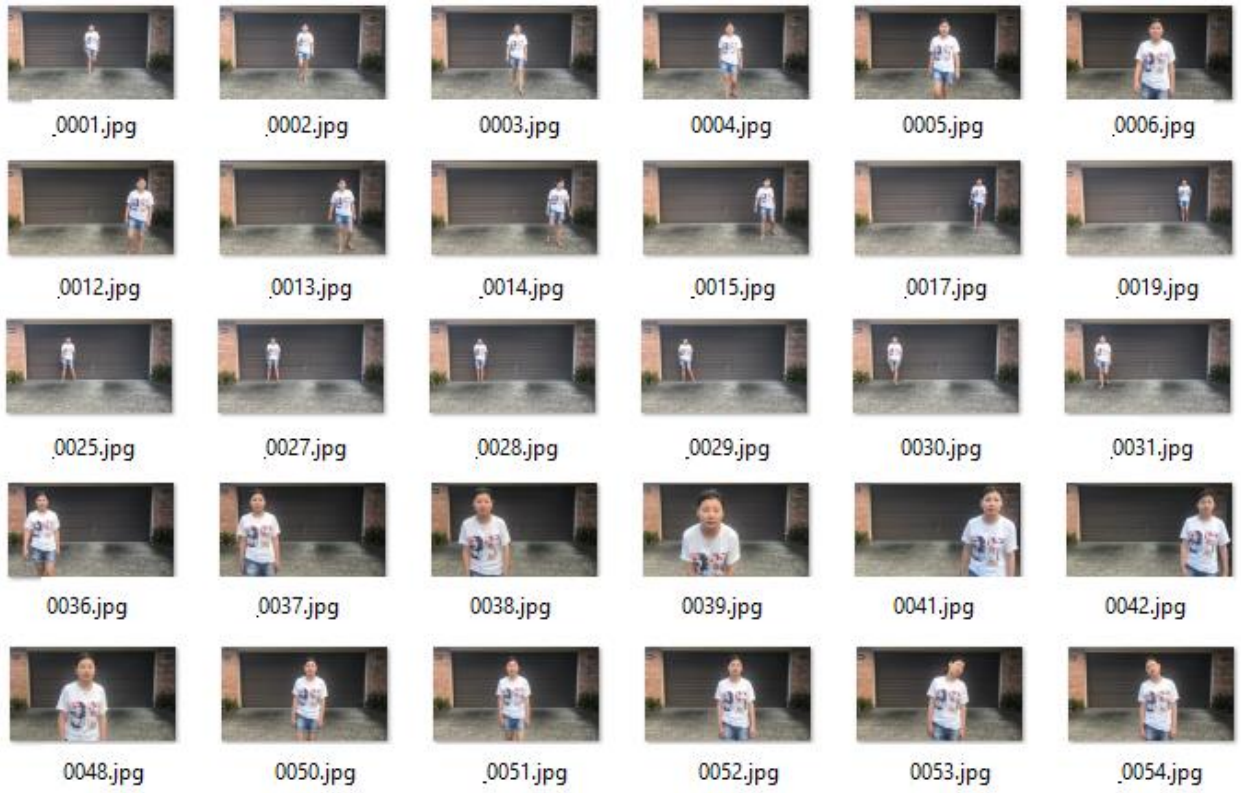

Figure 1 A part of the training dataset

In the following convolutional layers of this model, the dimension represents the size of images that has height *H* and width *W*, channel number is *C*. Then, six convolutional layers were setup in this model. In these six convolutional layers, a 5×5 large convolution kernel was setup in the first convolutional layer to capture a wide range of information and assign a 3×3 small convolution kernel at the next five convolutional layers, which is applied to capture a small range of image information. In each of the six convolutional layers, a 2×2 maximum pooling layer is set for each convolutional layer. The maximum pooling operation means that for each 2×2 small mesh, the maxima of pixel intensity has been taken as the output, which reduces the size of the special name at the same time.

After feature extraction of six convolutional layers, the feature maps of each sample were organized into a vector and further converted to a matrix. By expanding the vectors, dropout operations were added. The dropout operation indicates that at the time of output, the corresponding nodes are randomly deleted. If the dropout is conducted before fully connected layers, overfitting could be prevented effectively.

Then, taken the extracted visual features of CNN as input, two dual-layer MLPs were constructed respectively for the classification and localization of visual object. The dimension of the output of the classified MLP is [*N*, *nClass*], where *nClass* represents the number of classes. In our experiment, *nClass* = 5. Each row of the output represents a classification vector,

$$p = [p_1, \cdots, p_{nClass}], \tag{3}$$

where p_k represents the probability of the k-th classification of the input, which is applied to measure the gap between the predicted classification and the actual classification of the sample, cross entropy is generally employed as a loss function. The predicted classification vector of the given inpout is

$$q = [q_1, \cdots, q_{nClass}]. \tag{4}$$

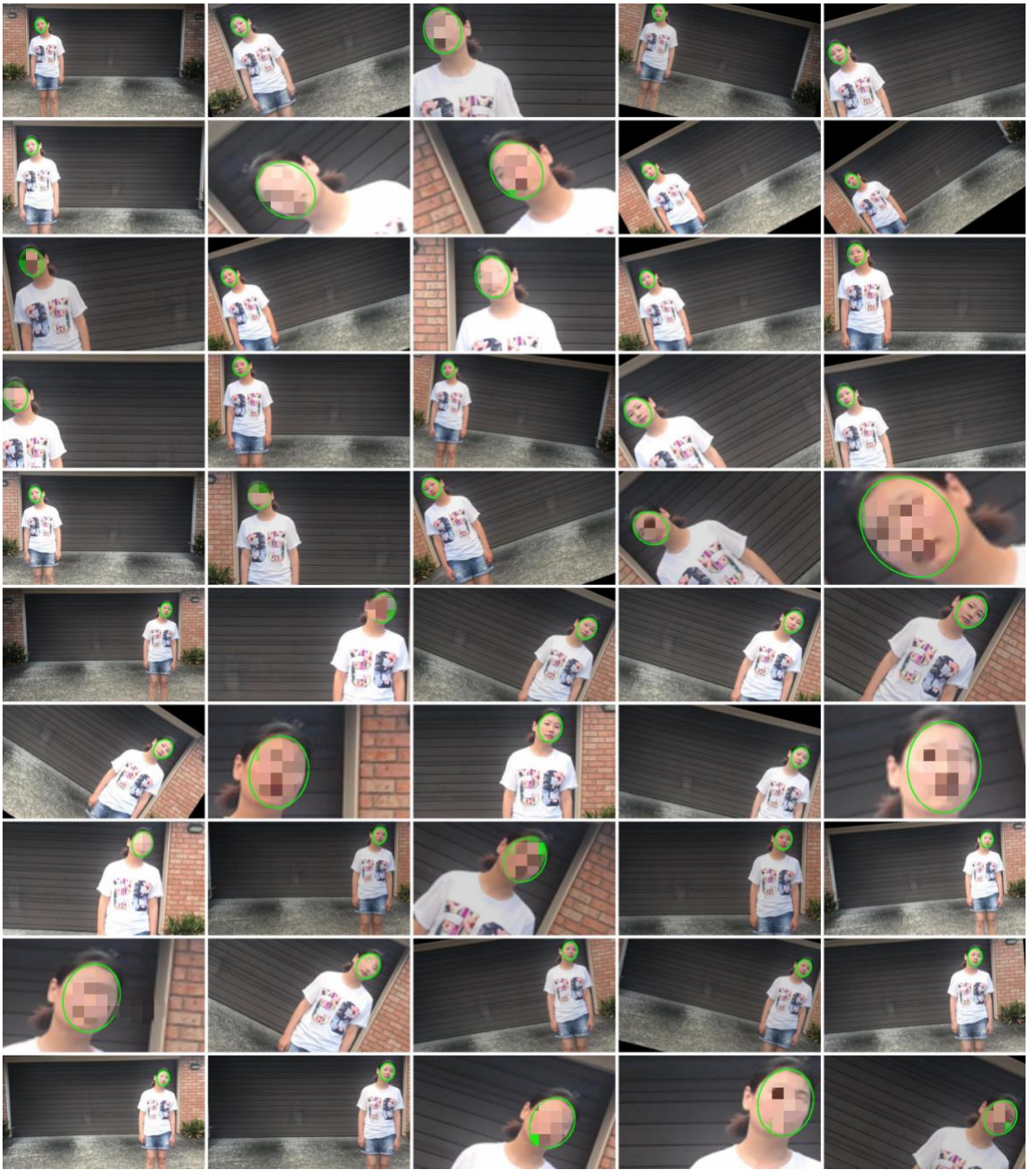

Figure 2 Images after randomly cropping

It satisfies that

$$q_k = \begin{cases} 1 & k = j \\ 0 & k \neq j \end{cases}. \tag{5}$$

The loss function of classification model is

$$L_{cls} = -\sum_{k=1}^{nClass} q_k \, log \, p_k \tag{6}$$

Because the feature map of location MLP is [N, 5], each row of the output result represents the code of an ellipse shape. Let $\hat{E}$ be the ellipse position predicted by the model, E be the actual ellipse position of the sample, then the location loss function of model is

$$L_{loc} = \sum_{k=1}^{5} f\left(\hat{E}_k - E_k\right). \tag{7}$$

The norm of vector $\hat{E}_k - E_k$ was calculated. The smooth 1-norm of function $f(\cdot)$ is chosen, then

$$f(x) = \begin{cases} |x| - 0.5 & x < 1 \\ 0.5x^2 & x \geq 1 \end{cases} \tag{8}$$

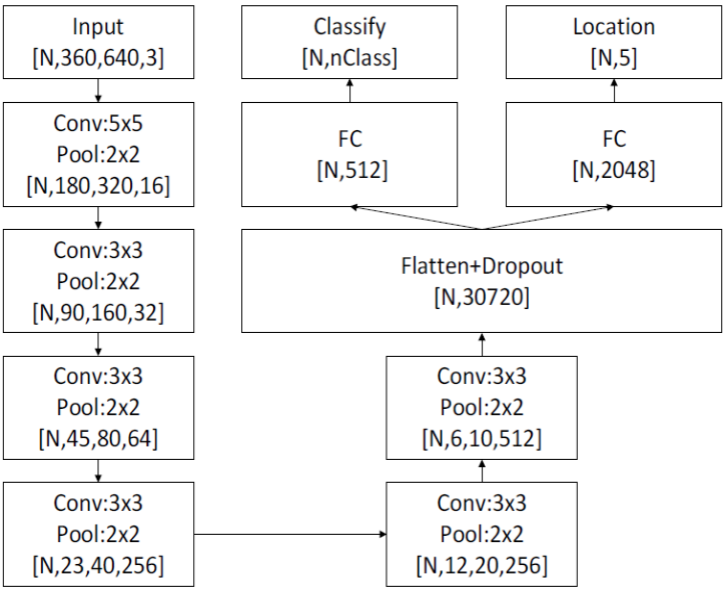

Figure 3 The structure of the proposed model

RESULT AND ANALYSIS

Corresponding to the research objectives of this book chapter, three experiments were conducted. In order to complete the task of currency detection, the secondary work embarked on is the stages of research work from data generating, training data and resultant analysis. The experimental implementation can considerably benefit from the basic idea regarding the specific process of currency identification.

According to the model design, the two sets of codes $E_1 = [x_c, y_c, a, b, 180\phi/\pi]$ and $E_2 = [x_1, x_2, y_1, y_2, b]$ are taken for model training. By taking weight $w \in \{1,10\}$, a total of four models were trained. Related to 2,5000 training samples, a batch size of 50 was taken with 75 epochs, which requires a total of 30,000 steps. The model was implemented by using TensorFlow and NVIDIA GTX 1070 graphics card for accelerated training, each model training process spent about 4 hours. The training results are shown in Table 1.

Table 1. Training results comparation

Models	Bounding Boxes	Validations
I	$[x_c, y_c, a, b, 180\phi/\pi]$	0.8890
II	$[x_c, y_c, a, b, 180\phi/\pi]$	0.8832
III	$[x_1, x_2, y_1, y_2, b]$	0.8962
IV	$[x_1, x_2, y_1, y_2, b]$	0.9018

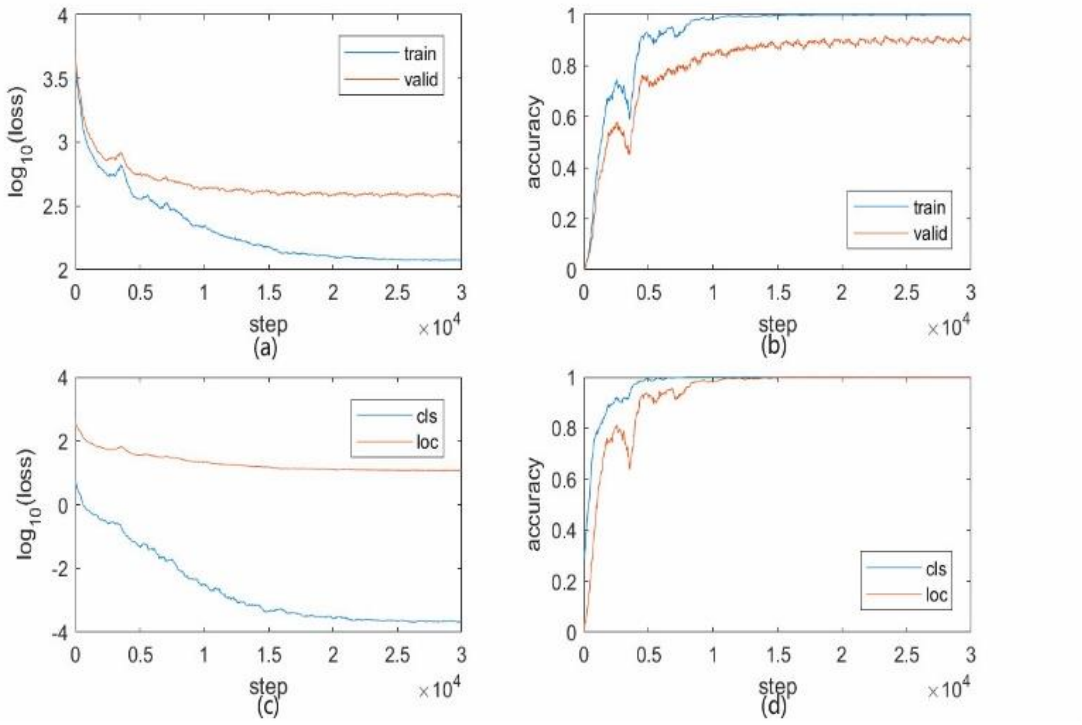

Figure 4 The curve of training process

The experimental results show that the accuracy of each model in the training is more than 99.00%, which indicates that the model has been fully trained. The correctness of the model is mainly compared by using

the accuracy based on the verification set. Table 1 shows the model IV with elliptic code E2, positioning weight w=10 has the highest accuracy based on the verification set, which reached to 90.18%. In order to further understand the training process of our model, a training curve is shown in Fig. 4.

Figure 4(a) represents the loss function value, which was mainly employed to observe the overfitting of the model. If the loss function on the training set decreases while the loss function on the validation set does not increase, then the model did not appear overfitting. Figure 4(a) shows the model had not apparently overfitting.

Figure 4(b) shows the results to compare the accuracy of the training and verification to observe whether the model has been adequately trained. The model basically converges at about 20,000 steps, the accuracy on the training set is almost 100.00%, which indicates that the model has been adequately trained.

Figure 4(c) shows the comparison of the loss function values located and classified on the training set. The Figure 4(d) displays the accuracy for positioning and classification on the training set. The decrease of the value of the classification loss function is larger than the value of the loss function of localization.

Of course, various sizes of a face will be obtained at different distance. As shown in Figure 5, if a person moves closer to the camera, deep learning algorithms are able to detect the human face with a diversity of sizes in different distance. The size of the face in the image is an important factor that affects the recognition results.

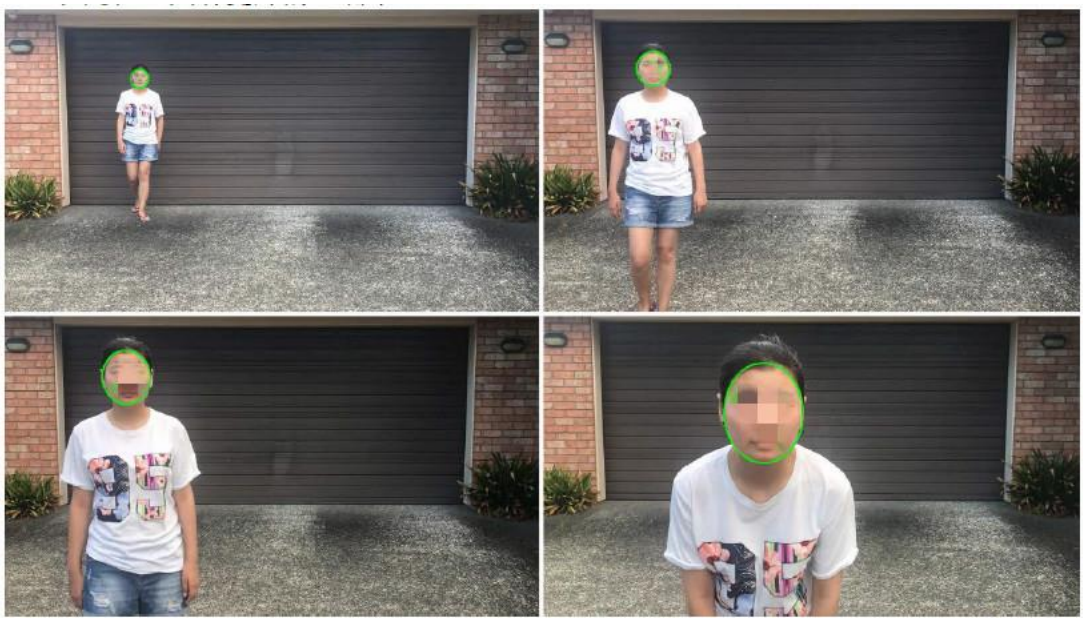

Figure 5 Human face with different sizes

The size of human face is measured by using proportion of the face to the total size of the image. Let the width and height of the image be w and h, respectively; the ellipse shape of a face is denoted as $\{x_c, y_c, a, b, \phi\}$, then the area of the bounding box of the human face is

$$S_F = 4\sqrt{(a\cos\phi)^2 + (b\sin\phi)^2}\sqrt{(a\sin\phi)^2 + (b\cos\phi)^2} \quad (9)$$

Thus, the ratio of the face in the image is obtained by using

$$A = \frac{4}{wh}\sqrt{(a\cos\phi)^2 + (b\sin\phi)^2}\sqrt{(a\sin\phi)^2 + (b\cos\phi)^2} \quad (10)$$

Obviously, if *A* is larger, it indicates a larger face in the image, people are closer to the camera; otherwise, if *A* is smaller, it reveals a smaller face is detected in the image, that represents that human face is far from the camera. In order to better understand the influence of proportion of a human face on the accuracy, the accuracy rates and IoUs for different sizes of faces were calculated.

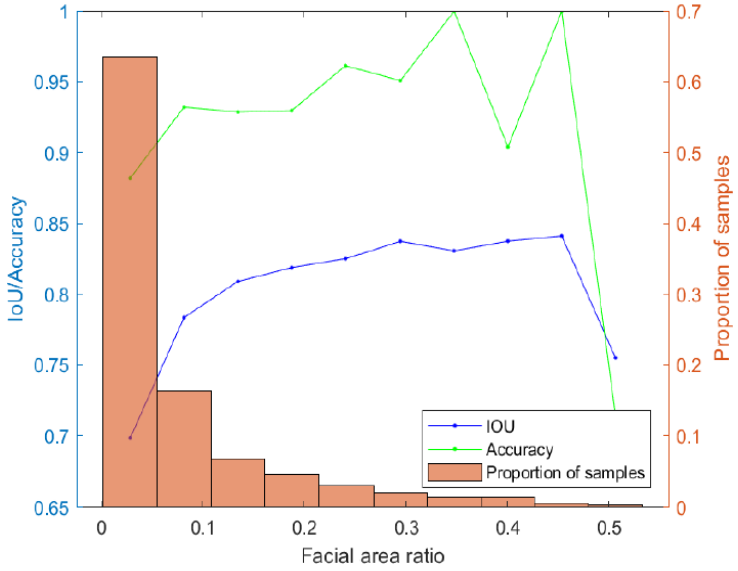

Figure 6 The relationship between face ratio and accuracy

From the histogram, if the proportion of a human face is small, the number of samples is large; if the ratio of faces increases, the number of samples gradually decreases. At the same time, both the accuracy and the IoU are increasing at first and then decrease. The reasons are: (1) If the face occupies a large proportion of the image, it is easily to be recognized and the accuracy is high; (2) The proportion of the face area to the image is large, the number of samples is less. Therefore, the values of accuracy and IoU are around the ratio 0.40, which begins to decline, because there are few samples, the statistical results fluctuate obviously.

CONCLUSION

There are two types of algorithms for detecting human faces from digital images and videos which are conventional machine learning and deep learning. Both algorithms are able to achieve high accuracy in face recognition, but conventional machine learning was not able to make a significant breakthrough based on aging, distance, partial occlusion, facial expression (Feng, Yuen, & Dai, 2000, Cui, 2014, Cui & Yan, 2016). In this book chapter, the use of ConvNets algorithms is an excellent solution on human face recognition, which greatly improves the efficiency of face recognition. The test videos contain a face with a broad range

of distances, the proposed model can still accurately detect the human face. Face detection using deep learning algorithms is beyond the recognition of using conventional machine learning methods.

Deep learning algorithms are effectively applied to solve a variety of external interferences in face recognition, but which requires a bit big dataset and huge training time. It may take much time in image augmentation and model training, even using a powerful GPU. In addition, the face detection algorithm by using deep learning algorithm in this book chapter mainly refers to the SSD model (Liu, *et al.*, 2016). There are many other deep learning algorithms that might optimize the neural network model, such as sparse coding, autoencoder, and deep belief network (Arel, Rose & Karnowski, 2010).

In future, the proposed network model is needed to be optimized based on the existing results (Yan, 2015, Yan, 2019, Yan, 2021). Firstly, the network model needs to effectively be designed by matching with facial features. Secondly, the training dataset should be reduced as much as possible to improve the training efficiency and availability (Wang, 2018).

REFERENCES

Arel, I., Rose, D. C., & Karnowski, T. P. (2010). Deep machine learning - A new frontier in artificial intelligence research. *IEEE Computational Intelligence*, 5(4), 13-18.

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8), 1798-1828.

Bian, Z. P., Hou, J., Chau, L. P., & Magnenat-Thalmann, N. (2016). Facial position and expression-based human–computer interface for persons with tetraplegia. *IEEE Journal of Biomedical and Health Informatics*, 20(3), 915-924.

Bonetto, M., Carrato, S., Fenu, G., Medvet, E., Mumolo, E., Pellegrino, F. A., & Ramponi, G. (2015). Image processing issues in a social assistive system for the blind. In *International Symposium on Image and Signal Processing and Analysis* (pp. 216-221).

Cireşan, D., Meier, U., Masci, J., & Schmidhuber, J. (2011). A committee of neural networks for traffic sign classification. In *International Joint Conference on Neural Networks* (IJCNN) (pp. 1918-1921).

Cui, W. (2014) *A Scheme of Human Face Recognition in Complex Environments*. Master's Thesis, Auckland University of Technology, New Zealand.

Cui, W., Yan, W. (2016) A scheme for face recognition in complex environments. *International Journal of Digital Crime and Forensics* (IJDCF) 8 (1), 26-36.

Feng, G. C., Yuen, P. C., & Dai, D. Q. (2000). Human face recognition using PCA on wavelet subband. *Journal of Electronic Imaging*, 9(2), 226-234.

Fukushima, K., & Miyake, S. (1982). Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In *Competition and Cooperation in Neural Nets* (pp. 267-285). Springer.

Gal, Y., & Ghahramani, Z. (2016). A theoretically grounded application of dropout in recurrent neural networks. In *Advances in Neural Information Processing Systems* (pp. 1019-1027).

Gao, X., Nguyen, M., Yan, W. (2021) Face image inpainting based on generative adversarial network. In *International Conference on Image and Vision Computing New Zealand*.

Hinton, G. E., Osindero, S., & Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7), 1527-1554.

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1725-1732).

Kingsbury, B. (2009). Lattice-based optimization of sequence classification criteria for neural-network acoustic modeling. In *IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 3761-3764).

Kong, T., Yao, A., Chen, Y., & Sun, F. (2016). Hypernet: Towards accurate region proposal generation and joint object detection. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 845-853).

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems* (pp. 1097-1105).

Landahl, H. D., McCulloch, W. S., & Pitts, W. (1943). A statistical consequence of the logical calculus of nervous nets. *Bulletin of Mathematical Biology*, 5(4), 135-137.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436.

Lee, Y., Kim, H., Park, E., Cui, X., & Kim, H. (2017). Wide-Residual-Inception networks for real-time object detection. In *IEEE Intelligent Vehicles Symposium* (pp. 758-764).

Lin, M., Chen, Q., & Yan, S. (2013). Network in network. arXiv:1312.4400.

Liu, J., & Liang, D. (2005). A survey of FPGA-based hardware implementation of ANNs. In *IEEE International Conference on Neural Networks and Brain* (pp. 915-918).

Liu, M., Yan, W. (2022) Masked face recognition in real-time using MobileNetV2. *ACM ICCCV*.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016). SSD: Single shot multibox detector. In *European Conference on Computer Vision* (pp. 21-37).

Maalej, R., Tagougui, N., & Kherallah, M. (2016). Online Arabic handwriting recognition with dropout applied in deep recurrent neural networks. In *IAPR Workshop on Document Analysis Systems* (pp. 417-421).

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 5(4), 115-133.

Nagi, J., Ducatelle, F., Di Caro, G. A., Cireşan, D., Meier, U., Giusti, A., ... & Gambardella, L. M. (2011). Max-pooling convolutional neural networks for vision-based hand gesture recognition. In *IEEE International Conference on Signal and Image Processing Applications* (pp. 342-347).

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, 61, 85-117.

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1), 1929-1958.

Wang, H. (2018), *Real-time Face Detection and Recognition Based on Deep Learning*. Master's Thesis, Auckland University of technology, New Zealand.

Wang, Z., & Srinivasan, R. S. (2017). A review of artificial intelligence-based building energy use prediction: Contrasting the capabilities of single and ensemble prediction models. *Renewable and Sustainable Energy Reviews*, 75, 796-808.

Wu, M. N., Lin, C. C., & Chang, C. C. (2007). Brain tumor detection using color-based k-means clustering segmentation. In *International Conference on Intelligent Information Hiding and Multimedia Signal Processing* (pp. 245-250).

Wu, J., Leng, C., Wang, Y., Hu, Q., & Cheng, J. (2016). Quantized convolutional neural networks for mobile devices. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4820-4828).

Yan, W., Kankanhalli, M. (2015) Face search in encrypted domain. Pacific-Rim Symposium on Image and Video Technology, 775-790.

Yan, W. (2019) *Introduction to Intelligent Surveillance: Surveillance Data Capture, Transmission, and Analytics*. Springer.

Yan, W. (2021) *Computational Methods for Deep Learning: Theoretic, Practice and Applications*. Springer.

Zhang, S., Wei, Z., Nie, J., Huang, L., Wang, S., & Li, Z. (2017). A review on human activity recognition using vision-based method. *Journal of Healthcare Engineering*, 2017.