Efficient and Effective Multi-person and Multi-angle Face Recognition based on Deep CNN Architecture

An-Chao Tsai[1], Yang-Yen Ou[2], Liu-Yi-Cheng Hsu[2], Jhing-Fa Wang[1,2]
[1]Department of Digital Multimedia Design, Tajen University.
[2] Department of Electrical Engineering, National Cheng Kung University.
E-mail: actsai@tajen.edu.tw

***Abstract*—Recently, the development and application of robots has become a famous topic and the most important thing for robots is personification. This paper uses a webcam to capture the image as a visual system input. The facial image is obtained through high performance face detect neural network. Facial landmarks are used to correct the face. Then, we use facial color RGB images for facial feature detection and identity recognition. By training a complete feature detection network, it is possible to detect valid and distinct facial features and train the classifier for those features. We can obtain identity confidence by using classifier for those feature. The experiment results show that the accuracy of identity recognition can be as high as 90.61%. In practical applications, the system can recognize identities up to thousands of people at the same time.**

Keywords—Multi-person and Multi-angle Face Recognition, Deep Convolutional Neural Network.

I. INTRODUCTION

Human face recognition system has been development for nearly 30 years [1]. At present, face recognition [2] technology has been introduced into 40 countries around the world, with more than 100 systems. In 2014, companies including Google, Baidu IDL, Microsoft and Facebook, as well as startups such as Ginger and Face++ [3], gained great attention because of their research on artificial intelligence such as deep learning and machine vision [4, 5].

The motivation of the proposed work focused on improving the image understanding about human identity. This work provided an automatic understanding system with deep convolution neural network for identity recognition system [6, 7]. The training concept is different from the previous methods, robots need to sense information from the outside world to decide their own behavior and interact with people. The robot is no longer a machine. A robotic system that understands human vision can be used in a wide range of situations, such as customer identification, surveillance and health-care, etc. This work try to understand human identities by using the facial image with deep learning architecture. We define the identity and emotion of the face in the screen.

The organization of the paper is as follows. In Section II, we described the related work for our research. Section III shows the proposed system. Experiment result is shown in Section IV. Finally, the conclusion is given in Section V.

II. RELATED WORK

A. Deep Convolutional Neural Network

The deep architectures [3, 8, 9] have been developed for some time [6] but there was no successful nor efficient methods reported for image recognition until 1989 because of the difficulty in training deep networks. Therefore, in order to deal with the problems that fully-connected layers faced when processing images, the Convolutional Neural Network (CNN) was firstly introduced in Computer Vision for image recognition by LeCun et al. in 1989 [10]. CNNs has been widely use in the 90s of 20th century but is gradually decreased with the advent of SVM [5] and Bayesian models. With the advent of larger datasets such as ImageNet in the first decade of 21st century, training in deep convolutional neural networks has become feasible.

The AlexNet was proposed by Krizhevsky et al. [11] in 2012 for ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) competition. Krizhevsky successfully demonstrate the efficiency of data augmentation and the use of recitified linear activation function. In 2013, the model named ZFNet, which was modified by AlexNet, built by Zeiler et al., won the ILSVRC competition. The VGGNet is an improvement of AlexNet in 2014 which also performs well in extracting features from images. The structure of AlexNet is deeper than that of AlexNet which applied with smaller filters and is proved to be more efficient. On the other hand, Szegedy et al [12] proposed GoogleNet (Inception V1), a neural network architecture that concatenate and merge different size of filters together to effectively reduce computing resources [12] and is the winner of ILSVRS 2014. At the end of 2015, Microsoft Research team has developed ResNet [4] with lower complexity than VGGNET but with a better performance. With the advanced of the hardware improvement, especially in graphic processing unit (GPU), the Convolutional Neural Network has been wildly recognized as a new solution in image and video processing. Therefore, how a develop an efficient and effective CNN architecture to achieve better performance would be our first concern.

B. Survey of Face Recognition

Great improvement has been achieved in facial recognition over the past few years. More and more systems attain excellent performance, even in the unconstrained environments is now no longer an issue with the support of hardware and algorithms. The famous Labeled Faces in the Wild (LFW) [13] data set includes facial images with different kind of variations in pose, facial expression and illumination which is normally used and compared as a benchmark for the research of facial recognition systems. With 97.35% recognition accuracy, Facebook's DeepFace was the first system to achieve near human-level performance (97.53%) [14] on this data set [15]. After the facial alignment, they use a 9-layer deep neural network with more than 120 million parameters in the system. These parameters were learned from a labelled data set containing four million images of over 4,000 identities. As a result, they received a face representation that generalizes well in comparison with other

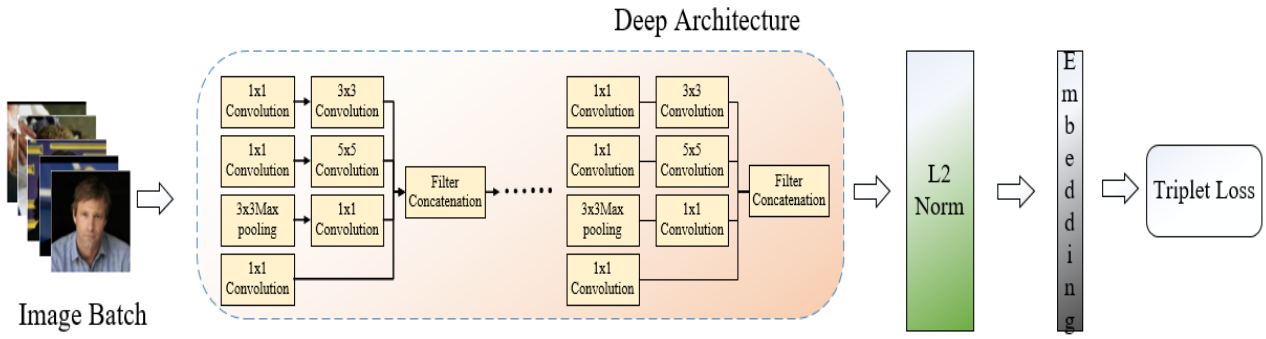

Figure 1. Neuro Network Architecture of Multi-person and Multi-angle Face Recognition

data sets such as LFW. Gaussian Face was the first system that superior human-level performance with an accuracy of 98.52% on LFW [13].

III. Proposed Method

The proposed research is based on Facial RGB image [2, 16] as input of visual system [17]. We first input the facial image into a deep CNN architecture, the deep architecture is used to remove the structure of softmax from deep convolutional neural network. In this paper, we adopted Google Inception V3 as our deep learning architecture. When the image pass through the deep architecture as well as the L2 normalization, the feature vectors which embedding into Euclidean space will be collected (each facial image generates a feature vector). Once the Euclidean space is created, the remaining tasks for face verification, face recognition and face clustering will become straightforward. Finally, the classifier of these vectors would be trained to perform face recognition. The proposed architecture is shown as Figure 1.

A. Deep Architecture for feature extraction

The google inception v3 is selected as our feature extraction [18] network since it is different from the previous neural network nor some common methods. This network reduces huge computational complexity required by deep convolutional neural networks. The network of Inception v3 are with a 1×7, 7×1, 3×1 and 1×3 convolution kernel to speed up the operation and a 1×1 convolution kernel (Network in Network) to further reduce parameters. While ignoring the parameters, the network speeds up the computing time and capable to make instant identification. In the meanwhile, a layer of nonlinear extended model is added to the network. The network structure is shown as Figure 2.

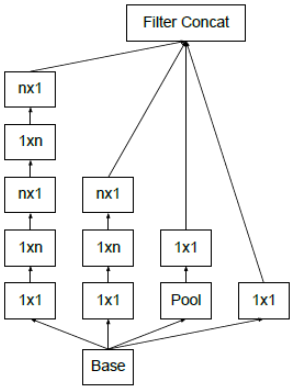

Figure 2. Google Inception V3

Instead of using the entire google Inception v3, the softmax layer has been removed to fit the target application. Since the database may gradually increase or decrease, removing softmax layer can avoid inequality of identity data. If not, to retrain the back end fully connected network will be required, which is a complicated and time consuming process, especially for environments without GPU. The proposed work becomes a training method using the triplet which is enough to represent a person with 128-dimensional feature vectors.

B. L2 normalization

Totally, 128-dimension feature vectors from the google inception network has been extracted. In order to map the feature vectors to a d-dimension hypersphere, i.e. $\|f(x)\|_2 = 1$. We applied L2-norm to normalize the feature vectors. The difference between two identities can be known by calculating the distance between two vectors in Euclidean space and can be represented as follows:

$$\boldsymbol{x} = [\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots., \boldsymbol{x_d}]$$

$$\boldsymbol{y} = [\boldsymbol{y_1}, \boldsymbol{y_2}, \ldots., \boldsymbol{y_d}]$$

$$\boldsymbol{y} = \frac{\boldsymbol{x}}{\sqrt{\sum_{i=1}^{d} \boldsymbol{x_i^2}}} = \frac{\boldsymbol{x}}{\sqrt{\boldsymbol{x^T x}}}$$

Where $x = [x_1, x_2, \ldots, x_d]$is sample input with dimension d, $y = [y_1, y_2, \ldots, y_d]$ is the forward pass output. For a vector (sample), its L2 norm can be derived as $\sqrt{\sum x_i^2} = \sqrt{x^T x}$.

C. Feature vectors for classification

Each face of every identity will attain its corresponding feature vectors in 128-dimension. We then build the Euclidean space, face recognition, verification and clustering. The Euclidean space can be used for classification between testing image and database but we still need to train a classifier to increase the recognition rate.

The Mulit-class SVM is chosen as our classifier which can be found in library named sklearn in python. This library includes Multi-class SVM which uses the concept of one-versus-one. The idea is to design an SVM between any two types of samples. One-versus-one classification can be represented as Figure 3.

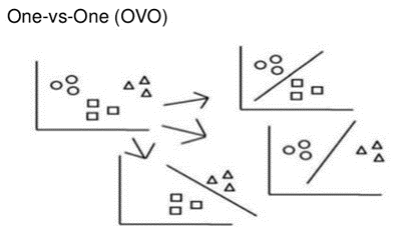

Figure 3. One-vs-One classification

When an unknown sample is classified, the category with the most votes is the category of the unknown sample. Finally, the embedding arrays and labels are inputted to train the classifier.

D. Training Phase

1) Triplet Loss

As mentioned above, the purpose of the model is to embed the face image into *d-dimensional* Euclidean space. The embedding can be represented by $f(x) \in R^d$. Otherwise, it constrains this embedding to live on the *d-dimensional* hypersphere, *i.e.* $\|f(x)\|_2 = 1$. In this space, we hope to ensure the image $x_i^a(anchor)$ of a specific person is closer to all other images $x_i^p(positive)$ of the same person than it is to any image $x_i^n(negtive)$ of any other person.

$$\left\|f(x_i^a) - f(x_i^p)\right\|_2^2 + \alpha < \|f(x_i^a) - f(x_i^n)\|_2^2$$

$$\forall(f(x_i^a), f(x_i^p), f(x_i^n)) \in \tau$$

Where α depicts a margin that is enforced between positive and negative pairs. τ is the set of all possible triplets in the training set with cardinality N. Therefore, to rewrite the above equation, the loss function is derived as:

$$\sum_i^N \left[\left\|f(x_i^a) - f(x_i^p)\right\|_2^2 - \|f(x_i^a) - f(x_i^n)\|_2^2 + \alpha\right]_+ .$$

2) Triplet Selection

For each x_i^a in the training sample, we need to select the different image of same individual x_i^p so that its Euclidean distance is the furthest. Conversely, the nearest should also be found, both of them are defined as follows:

$$argmax_{x_i^p}\left\|f(x_i^a) - f(x_i^p)\right\|_2^2$$

$$argmin_{x_i^n}\|f(x_i^a) - f(x_i^n)\|_2^2$$

In actual training, it is impractical to calculate all $argmax$ and $argmin$ in the dataset, because the amount of computational complexity. Moreover, in database, there will be some wrong labels which may be combined into triplets and lead to training difficulties. With these observation we proposed a triplet selection mechanism as follows:

> "Generate Triplets online and select the extreme pos/neg sample in each mini-batch."

In order to confirm the triplet would be generated in mini-batch, we ensure each person in each mini-batch has an average of 40 images when generating mini-batch. And randomly select other face pictures as negative samples. Improper selection of negative samples may lead local minimums in training. To avoid this situation, we use the following formula to filter out negative samples:

$$\left\|f(x_i^a) - f(x_i^p)\right\|_2^2 < \|f(x_i^a) - f(x_i^n)\|_2^2$$

3) Training Setting

The CNN has been trained using Stochastic Gradient Descent (SGD) with Backpropagation (BP) and AdaGrad. The models are randomly initialized. We start to train models with a learning rate of 0.05 and reducing the learning rate gradually. Models are trained in the environment with GPU (1080Ti) for about 160 hours. The margin α is set to 0.2.

IV. EXPERMENT RESULT

In this work, a RGB image with User Identity Dataset (UID) is proposed as a preliminary work for simulating the customer's daily use. Part of the collected dataset images are provided in Figure 4.

Since the camera may not capture the customer and user in a frontal face, to deal with this, multiple person and multiple angle are considered in our database. According to the multi-angle facial image, it makes the system more robust to the environmental and practical application uncertainties. In order to further increase the robust of the different scale of the face, we choose to build multiple database of different sizes depends on the distance from the camera.

The chosen camera is a commercial webcam. To simulate the height of the general robot or human eye, we set the camera about 1-meter height from the ground. Color image are 800x600 pixels in resolution. According to the actual test, we found that the longest distance is approximately 4 meters from the camera. Therefore, the database is collected with different distance from 1 to 3 meter. For multiple angle, set the face angle between $+75^{\circ}\sim-75^{\circ}$ for yaw and $+15^{\circ}\sim-15^{\circ}$ for pitch and took three photos for every five degrees except degree 0. As a result, the dataset contains 4,104 images collected from 12 participants who are master students in our lab and act as a user of 12 people. The experiments are conducted with different distance (1M, 2M and 3M) in real time testing.

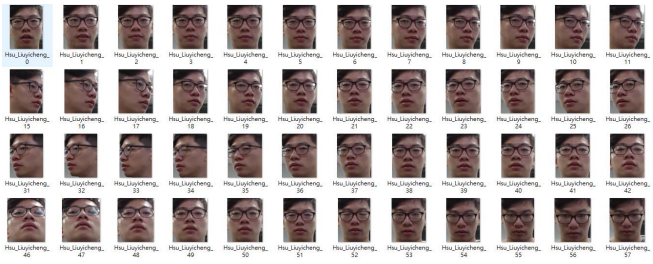

Figure 4. The RGB image with User Identity Dataset.

Table1 The Online Test Result of 1M for identity recognition

	Frontal	Yaw	Pitch	Average
User1	100%(10/10)	90%(9/10)	100%(10/10)	96.67%(29/30)
User2	100%(10/10)	60%(6/10)	100%(10/10)	86.67%(26/30)
User3	100%(10/10)	90%(9/10)	100%(10/10)	96.67%(29/30)
User4	90%(9/10)	100%(10/10)	100%(10/10)	96.67%(29/30)
User5	90%(9/10)	100%(10/10)	90%(9/10)	93.33%(28/30)
User6	100%(10/10)	100%(10/10)	90%(9/10)	96.67%(29/30)
User7	90%(9/10)	90%(9/10)	100%(10/10)	93.33%(28/30)
User8	100%(10/10)	100%(10/10)	100%(10/10)	100%(30/30)
User9	100%(10/10)	90%(9/10)	100%(10/10)	96.67%(29/30)
User10	90%(9/10)	80%(8/10)	100%(10/10)	90%(27/30)
User11	100%(10/10)	90%(9/10)	90%(9/10)	93.33%(28/30)
Stranger	60%(6/10)	60%(6/10)	70%(7/10)	63.33%(19/30)
Total	93.33 (112/120)	87.5(105/120)	95%(114/120)	**91.94(331/360)**

Table2 The Online Test Result of 2M for identity recognition

	Frontal	Yaw	Pitch	Average
User1	90%(9/10)	100%(10/10)	100%(10/10)	96.67%(29/30)
User2	100%(10/10)	90%(9/10)	100%(10/10)	96.67%(29/30)
User3	100%(10/10)	80%(8/10)	100%(10/10)	93.33%(28/30)
User4	90%(9/10)	100%(10/10)	100%(10/10)	96.67%(29/30)
User5	80%(8/10)	90%(9/10)	90%(9/10)	86.67%(26/30)
User6	100%(10/10)	90%(9/10)	80%(8/10)	90%(27/30)
User7	90%(9/10)	80%(8/10)	90%(9/10)	86.67%(26/30)
User8	100%(10/10)	100%(10/10)	100%(10/10)	100%(30/30)
User9	100%(10/10)	100%(10/10)	90%(9/10)	96.67%(29/30)
User10	100%(10/10)	70%(7/10)	100%(10/10)	90%(29/30)
User11	100%(10/10)	90%(9/10)	100%(10/10)	96.67%(29/30)
Stranger	90%(9/10)	90%(9/10)	90%(9/10)	90%(27/30)
Total	95%(114/120)	90%(108/120)	95%(114/120)	**93.33%(336/360)**

Table3 The Online Test Result of 3M for identity recognition

	Frontal	Yaw	Pitch	Average
User1	90%(9/10)	90%(9/10)	90%(9/10)	90%(27/30)
User2	100%(10/10)	80%(8/10)	100%(10/10)	93.33%(28/30)
User3	90%(9/10)	100%(10/10)	90%(9/10)	93.33%(28/30)
User4	80%(8/10)	80%(8/10)	100%(10/10)	86.67%(26/30)
User5	100%(10/10)	80%(8/10)	100%(10/10)	93.33%(28/30)
User6	90%(9/10)	100%(10/10)	80%(8/10)	90%(27/30)
User7	90%(9/10)	60%(6/10)	60%(6/10)	70%(21/30)
User8	100%(10/10)	90%(9/10)	80%(8/10)	90%(27/30)
User9	80%(8/10)	90%(9/10)	70%(7/10)	80%(24/30)
User10	100%(10/10)	90%(9/10)	90%(9/10)	93.33%(28/30)
User11	90%(9/10)	100%(10/10)	100%(10/10)	96.67%(29/30)
Stranger	100%(10/10)	90%(9/10)	100%(10/10)	96.67%(29/30)
Total	92.5%(111/120)	87.5%(105/120)	88.33%(106/120)	**89.44%(322/360)**

V. CONCLUSION

This work proposed a real time understanding system for face recognition with multi-angle and multi-person based on deep convolutional neural network. The different training concept for the identity network makes the classification and training of faces easier and faster. The experimental results have demonstrated the effectiveness of the proposed system, the recognition rate can achieve 90.61% for identity.

REFERENCES

[1] T. Mita, T. Kaneko and O. Hori, "Joint haar-like features for face detection," *Tenth IEEE International Conference on Computer Vision*, 2005, pp. 1619-1626.

[2] P. Viola and M. J. Jones, "Robust real-time face detection," *International journal of computer vision*, vol. 57, no. 2, pp. 137-154, 2004.

[3] G. Levi and T. Hassner, "Emotion recognition in the wild via convolutional neural networks and mapped binary patterns," *in Proceedings of the ACM on international conference on multimodal interaction*, pp. 503-510, 2015.

[4] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770-778.

[5] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt and B. Scholkopf, "Support vector machines," *IEEE Intelligent Systems and their applications*, vol. 13, no. 4, pp. 18-28, 1998.

[6] F. Zhi-Peng, Z. Yan-Ning and H. Hai-Yan, "Survey of deep learning in face recognition," *IEEE International Conference on Orange Technologies*, 2014, pp. 5-8.

[7] X. Cao, D. Wipf, F. Wen, G. Duan and J. Sun, "A Practical Transfer Learning Algorithm for Face Verification," *IEEE International Conference on Computer Vision*, 2013, pp. 3208-3215.

[8] P. Sermanet, S. Chintala, and Y. LeCun, "Convolutional neural networks applied to house numbers digit classification," *International Conference on Pattern Recognition (ICPR2012)*, 2012, pp. 3288-3291.

[9] A. T. Lopes, E. de Aguiar, A. F. De Souza and T. Oliveira-Santos, "Facial expression recognition with convolutional neural networks: coping with few data and the training sample order," *Pattern Recognition*, vol. 61, pp. 610-628, 2017.

[10] Y. LeCun et al., "Backpropagation Applied to Handwritten Zip Code Recognition," *Neural Computation*, vol. 1, no. 4, pp. 541-551, 1989.

[11] A. Krizhevsky, I. Sutskever and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, 2012, pp. 1097-1105.

[12] C. Szegedy et al., "Going deeper with convolutions," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1-9.

[13] G. B. Huang, M. Ramesh, T. Berg and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," *Technical Report 07-49, University of Massachusetts*, Amherst, 2007.

[14] N. Kumar, A. C. Berg, P. N. Belhumeur and S. K. Nayar, "Attribute and simile classifiers for face verification," *IEEE International Conference on Computer Vision*, 2009, pp. 365-372.

[15] Y. Taigman, M. Yang, M. A. Ranzato and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," *IEEE conference on computer vision and pattern recognition*, 2014, pp. 1701-1708.

[16] K. Zhang, Z. Zhang, Z. Li and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499-1503, 2016.

[17] B. Amos, B. Ludwiczuk and M. Satyanarayanan, "Openface: A general-purpose face recognition library with mobile applications," *CMU School of Computer Science*, 2016.

[18] S. Shojaeilangari, W.-Y. Yau, K. Nandakumar, J. Li and E. K. Teoh, "Robust representation and recognition of facial emotions using extreme sparse learning," *IEEE Transactions on Image Processing*, vol. 24, no. 7, pp. 2140-2152, 2015.