Age and Gender Prediction from Face Images Using Convolutional Neural Network

Koichi Ito*, Hiroya Kawai*, Takehisa Okano* and Takafumi Aoki*
* Graduate School of Information Sciences, Tohoku University, Sendai, Japan
E-mail: ito@aoki.ecei.tohoku.ac.jp

***Abstract*—Attribute information such as age and gender improves the performance of face recognition. This paper proposes an age and gender prediction method from face images using convolutional neural network. Through a set of experiments using public face databases, we demonstrate that the proposed method exhibits the efficient performance on age and gender prediction compared with conventional methods.**

I. Introduction

Biometric authentication, which identifies a person using physiological or behavioral characteristics, has attracted much attention because of providing better security and more convenience than conventional approaches such as key, password and number [1]. Various biometric traits such as face, fingerprint, iris, signature, voice and gait can be used in person authentication. Among them, face recognition is the hottest research topic in biometrics and is in a huge demand, since person authentication using face is a natural way of human beings [2].

The general flow of face recognition consists of three steps: capture face images by a camera with visible or near-infrared illumination, extract features from face images and evaluate the similarity between features. The advantage of face recognition is its convenient image acquisition procedure compared with other biometric recognition such as iris and fingerprint recognition, since special equipment is not required to capture face images. On the other hand, the disadvantage of face recognition is that face images are dramatically changed by head pose variations, expression changes, aging, etc., resulting in low recognition accuracy.

Addressing the above problem, there is an approach of combining multiple minute traits extracted from an input image to improve the recognition accuracy of conventional methods. Such minute traits may not have discriminative information in person authentication, although they include personal information. Biometric recognition using minute traits is called *soft biometrics* as distinguished from general biometrics [1]. Traits used in soft biometrics are, for example, gender, age, ethnicity, hairstyle, hair color, accessory, etc. These are not always useful in person authentication, while a set of them can be complementarily used with general biometric recognition and can be taken under various condition of data acquisition compared with general biometric traits. Therefore, the combined use of face recognition and soft biometrics makes it possible to improve the recognition accuracy of face recognition. This paper focuses on age and gender prediction from face images for the purpose of enhancing the performance of face recognition systems.

There are some methods in age and gender prediction from face images [3], [4], [5], [6]. Among them, Yi et al. [4] and Rothe et al.[6] proposed age and gender prediction methods using Convolutional Neural Network (CNN). Rothe et al. [6] employed VGG-16 [7], which is a simple sequential CNN architecture, to estimate age and gender from face images. A lot of CNN architectures, whose performance are higher than VGG-16, are available now.

This paper proposes an age and gender prediction method using CNN. We explore the best CNN architecture for estimating age and gender from face images. We also introduce the Multi-Task Learning (MTL) to improve the accuracy of age and gender prediction. MTL is one of machine learning approaches to improve the performance by training multiple tasks simultaneously. The approach of MTL applying to deep learning is called Deep Multi-Task Learning (DMTL). The idea of DMTL is that the prediction accuracy is improved by training the network with sharing a part of network for each task and the loss function [8], [9], [10]. We explore the effectiveness of DMTL in age and gender prediction from face images.

Guo et al. [3], Yi et al. [4] and Han et al. [5] used the MORPH album2 dataset (MORPH II), while Rothe et al. [6] and Ricaneck et al. [11] used the IMDB-WIKI dataset. This paper uses the IMDB-WIKI dataset both in training and test because of its large size and its wide distribution of age. Note that data cleaning is required in advance, since there are a lot of images with a wrong label and inadequate images (i.e., non-face images) included in the IMDB-WIKI dataset.

II. Proposed Method

This section describes the proposed method using CNN for age and gender prediction. We explain the detail of the network architecture and the DMTL approach used in the proposed method in the following.

A. Network Architecture

The performance of CNN is directly influenced by the depth of network. Therefore, the recent network architectures such as GoogLeNet [12] and WideResNet [13] employ parallel network layers to exhibit the good performance of CNN. The large size of network exhibits the good performance on image

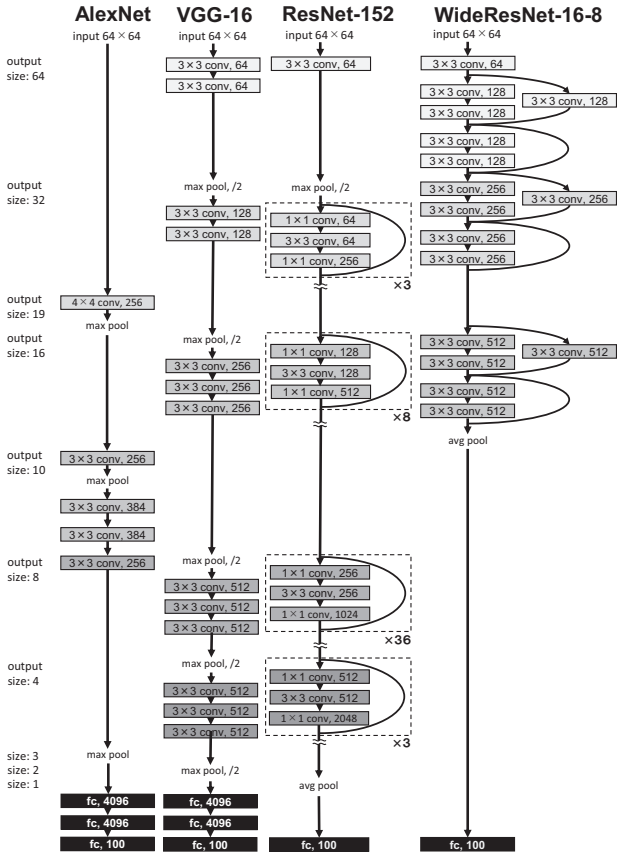

Fig. 1. Network architectures of CNN used in the proposed method.

recognition, while a huge amount of processing time or high-performance computers is required in training. Therefore, the optimal network architecture for CNN has to be selected taking into account the balance of computation time and hardware resources and also depending on applications. This paper explores four network architectures: AlexNet [14], VGG [7], Residual Network (ResNet) [15] and Wide Residual Network (WideResNet or WRN) [13], where each architecture won the first prize in the past ImageNet Large Scale Visual Recognition Competition (ILSVRC)[1].

Fig. 1 shows network architectures of CNN used in the proposed method. Note that the number of output layers is different depending on tasks. AlexNet [14], which won the first prize in ILSVRC 2012, employs Rectified Linear Units (ReLU) for the activation function and dropout for the fully-connected layers. Such fundamental techniques are commonly used even now. AlexNet consists of 5 convolution blocks and 3 fully-connected layers, which is the simplest and shallowest architecture in this paper. VGG [7], which won the first prize in ILSVRC 2014, employs deeper layers than AlexNet by using 3×3 convolution layers. This paper employs VGG-16, where the number of convolution layers is 16. ResNet [15], which won the first prize in ILSVRC 2015, introduces "shortcut connections" into the deep CNN architecture in order to control gradient loss, resulting in archiving deeper CNN architectures. This paper employs ResNet-152, where the number of convolution layers is 152. WideResNet [13] is one of extensions of ResNet, where the ResNet architecture is parallelized by adding a convolution layer to shortcut connections, where the number of filters in each convolution layer is controlled by the width parameter. The performance of WideResNet with 28 layers is higher than ResNet with 1,001 convolution layers. This paper employs WideResNet-16-8, where the number of convolution layers is 16 and the width parameter is 8.

[1] http://www.image-net.org/challenges/LSVRC/

B. Deep Multi-Task Learning

This paper employs multi-task learning (MTL) to improve the accuracy and reduce the computation time in age and gender prediction. The CNN architectures are modified to introduce MTL as shown in Fig. 2. MTL employs one feature extraction both for age prediction task and gender prediction task, while STL employs independent feature extraction for each task. Then, features output from the shared feature extraction are input to independent fully-connected layers for each task.

We also modify computation of the objective function. The minimization of objective function in STL is given by

$$\arg\min_{W} \sum_{i=0}^{N} \mathcal{L}(y_i, \mathcal{F}(X_i, W)) + \lambda\Phi(W), \tag{1}$$

where $\mathbf{X} = \{X_i\}_{i=0}^{N}$ and $\mathbf{Y}\{y_i\}_{i=0}^{N}$ are images and labels included in the training dataset, respectively, and W are weights. $\mathcal{F}$ is an estimation function using X_i and W, that is, neural network. $\mathcal{L}$ is a loss function and $\lambda\Phi(W)$ is a regularization term.

The minimization of object function in DMTL is given by

$$\begin{aligned}\arg\min_{W_s, W^a, W^g} \sum_{i=0}^{N} \{&\mathcal{L}^a(y_i^a, \mathcal{F}(X_i, W_s \circ W^a)) \\ &+ \mathcal{L}^g(y_i^g, \mathcal{F}(X_i, W_s \circ W^g)) \\ &+ \lambda_1\Phi(W_s) + \lambda_2\Phi(W^a) + \lambda_3\Phi(W^g)\},\end{aligned} \tag{2}$$

where W_s are shared weights between tasks, W^a and W^g are weights for age prediction and gender prediction tasks, respectively. y_i^a and y_i^g are age and gender of the image X_i, respectively. $\mathcal{L}^a$ and $\mathcal{L}^g$ are loss functions for age prediction and gender prediction, respectively. $\lambda_1\Phi(W_s)$, $\lambda_2\Phi(W^a)$ and $\lambda_3\Phi(W^g)$ are regularization terms for each weight. The whole network including shared parts is simultaneously trained by minimizing the objective function with loss function for each task.

III. Experiments and Discussion

This section describes performance evaluation of the proposed method using the IMDB-WIKI dataset.

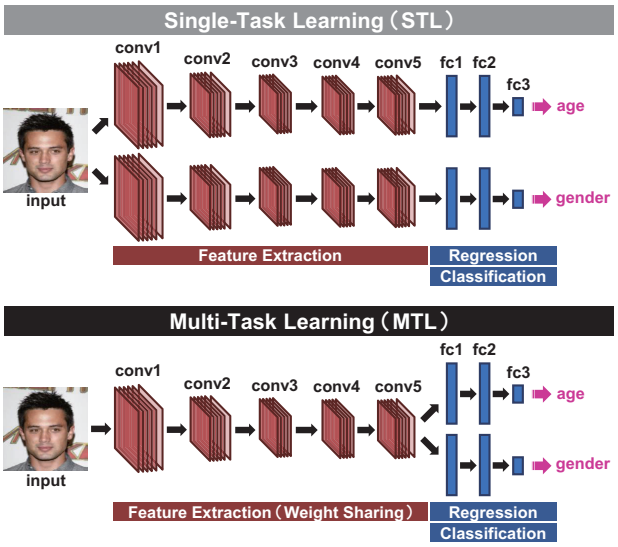

Fig. 2. Network architecture for STL and DMTL.

A. IMDB-WIKI Dataset

The IMDB-WIKI dataset consists of 523,051 images in IMDb[2] and Wikipedia[3]. This dataset includes the acquisition data, the date of birth, the gender, the face score of face detector and so on [16]. The age of subjects is calculated as the difference between the date of birth and the acquisition date, since the age label is not included in the dataset. This dataset includes some errors as shown in Fig. 3, for example, non-face images, images with wrong label, images with more than two faces and so on. Therefore, we must clean the IMDB-WIKI dataset in advance. We employ the following condition to remove such errors from the dataset:

- an age is under 0 or more than 101, where images with more than 101 years old are buildings,
- a gender is NaN (not a number),
- the maximum face score is less than 1.0 and
- the second maximum face score is not NaN.

After applying the above filtering, we get 209,990 face images with age and gender labels. 171,852 images from IMDb (called the IMDB dataset) are used in training and 38,138 images from Wikipedia (called the WIKI dataset) are used in test. We extract a face region from a face image using *frontal face detector* of Dlib[4], which is a Python library.

B. Experimental Condition

AlexNet, VGG, ResNet, WideResNet (WRN) are trained using the IMDB dataset for classification or regression. We use VGG-16 [7] whose number of layers is 16, ResNet-152 [15] whose number of layers is 152 and WideResNet-16-8 [13] whose depth is 16 and width is 8 as mentioned in the previous section.

[2]https://www.imdb.com/
[3]https://www.wikipedia.org/
[4]http://dlib.net/

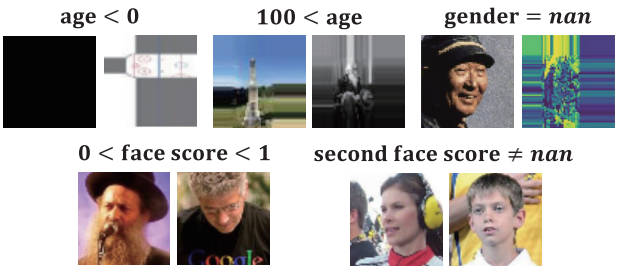

Fig. 3. Example of wrong images included in the IMDB-WIKI dataset.

An Input image is resized to 64×64 pixels and then its pixel values are normalized so as to have a mean 0 and a variance 1. 10% of the training dataset are used to check overfitting in validation of training. NAG (Nesterov Accelerated Gradient) is used in optimization of training and a loss for validation data (*val loss*) is calculated for each epoch. The weights having the lowest *val loss* are used in final. We evaluate the accuracy of gender estimation and the accuracy of mean absolute error (MAE) of age estimation to compare the performance of each method. Note that, in the case of regression, the accuracy is calculated using integer values obtained by rounding outputs while the MAE is calculated using floating-point values.

The processing time for each architecture is also evaluated. We evaluate the processing time for each architecture until the estimated age is output in STL and until the estimated age and gender are output in DMTL, when inputting one image. The configuration of computation environment is as follows:

- PC: Microsoft® Surface™ Pro 4,
- OS: Microsoft® Windows® 10 Pro,
- CPU: Intel® Core™ i5-6300U @ 2.40GHz,
- RAM: 8.00GB and
- TOOL: Microsoft® Visual Studio Code.

Note that the effectiveness of DMTL is evaluated by age estimation in regression and gender estimation in classification.

C. Comparison of Network Architectures

Table I shows a summary of experimental results for each network architecture. WRN exhibits the best performance both on age and gender prediction. The computation time of AlexNet is the shortest, since its network architecture is simple. The wide architecture of WRN demonstrates the effectiveness in face image analysis. VGG can be used in practical use, since VGG has a good balance between the accuracy and the computation time compared with others. For every architecture, the accuracy of age estimation by regression is higher than that by classification. We can observe that the regression analysis is suitable for estimating continuous values such as age. On the other hand, it is different from each architecture which analysis should be used in gender estimation. Hence, we have to select regression and classification depending on architectures.

Fig. 4. Example of age and gender prediction using WRN, where the value in brackets indicates the ground truth.

TABLE I
SUMMARY OF EXPERIMENTAL RESULTS

Network		Gender	Age	
Model	Analysis	Accuracy	MAE	Time [s]
AlexNet	Classification	91.53%	9.44	0.038
	Regression	90.89%	8.98	0.040
VGG-16	Classification	93.36%	8.15	0.107
	Regression	93.43%	7.77	0.101
ResNet-152	Classification	92.11%	9.25	0.527
	Regression	92.99%	7.82	0.534
WRN-16-8	Classification	93.57%	8.59	0.544
	Regression	93.25%	7.52	0.533

Fig. 4 shows examples of age and gender estimation using WRN. The value below the image indicates an age, M (Male) and F (Female) indicate a gender and the value in bracket indicates the ground truth. Two columns in the left show correct estimation results, while two columns in the right show wrong estimation results. Faces with wrong gender estimation look like the opposite gender due to their hairstyle and shape. It is difficult even for human to predict a gender of such images. Faces with wrong age estimation include images such as photos and paintings, whose age is difficult even for human to estimate. Therefore, our dataset cleaning is not enough to remove such images with wrong label.

D. Effectiveness of Deep Multi-Task Learning

We evaluate the effectiveness of DMTL using WRN, whose accuracy is the best in the proposed method as mentioned above. Table II shows a summary of estimation accuracy for STL and DMTL. The use of DMTL makes it possible to significantly reduce the computation time compared with that of STL. The accuracy of age estimation is improved by DMTL, while the accuracy of gender estimation is degraded. The objective function of DMTL is defined by simply adding the loss in age estimation and gender estimation as shown in Eq. (2) The loss function is different from age and gender estimation, since age is estimated by regression and gender is estimated by classification. The value of loss function is more than 7.0 in age estimation since the loss is calculated by MAE, while that is less than 1.0 in gender estimation since the loss is calculated by the cross-entropy loss. Hence, weights may be optimized only for age estimation in training. To address the above problem, we have to introduce the balance factor between age and gender prediction to the objective function. As a result, we can observe that the accuracy of age estimation is improved by simultaneously training age and gender compared with STL. The use of DMTL makes it possible to improve the performance on CNN models in terms of the accuracy and the computation time.

TABLE II
COMPARISON BETWEEN STL AND DMTL.

Method	Gender Accuracy	Age MAE	Time [s]
WRN + STL	93.86%	7.327	1.322
WRN+DMTL	93.54%	7.217	0.804

IV. CONCLUSION

This paper proposed an age and gender prediction method from face images using CNN. We explored CNN architectures, regression/classification and STL/DMTL though the paper. WideResNet exhibits the best performance on age and gender prediction compared with other architectures, where age is estimated by regression and gender is estimated by classification. The use of DMTL makes it possible to improve the performance of CNN in terms of the accuracy and the computation time. In future work, we will consider the powerful data cleaning method, and improve the accuracy of the proposed method by modifying the DMTL approach.

REFERENCES

[1] A.K. Jain, P. Flynn, and A.A. Ross, *Handbook of Biometrics*, Springer, 2008.
[2] S.Z. Li and A.K. Jain, *Handbook of Face Recognition*, Springer, 2011.
[3] G. Guo and G. Mu, "A framework for joint estimation of age, gender and ethnicity on a large database," *Image Vision Comput.*, vol. 32, no. 10, pp. 761–770, Oct. 2014.
[4] D. Yi, Z. Lei, and S.Z. Li, "Age estimation by multi-scale convolutional network," *Proc. Asian Conf. Computer Vision*, pp. 144–158, 2014.
[5] H. Han, C. Otto, X. Liu, and A.K. Jain, "Demographic estimation from face images: Human vs. machine performance," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 37, no. 6, pp. 1148–1161, June 2015.
[6] R. Rothe, R. Timofte, and L.V. Gool, "DEX: Deep expectation of apparent age from a single image," *Proc. Int'l Conf. Computer Vision Workshops*, pp. 252–257, Dec. 2015.
[7] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.
[8] M. Ehrlich, T.J. Shields, T. Almaev, and M.R. Amer, "Facial attributes classification using multi-task representation learning," *Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshops*, pp. 47–55, June 2016.
[9] E.M. Hand and R. Chellappa, "Attributes for improved attributes: A multi-task network utilizing implicit and explicit relationships for facial attribute classification," *Proc. the Thirty-First AAAI Conf. Artificial Intelligence*, pp. 4068–4074, Feb. 2017.
[10] H. Han, A.K. Jain, F. Wang, S. Shan, and X. Chen, "Heterogeneous face attribute estimation: A deep multi-task learning approach," *IEEE Trans. Pattern Analysis and Machine Intelligence*, (to be published).
[11] K. Ricanek and T. Tesafaye, "MORPH: A longitudinal image database of normal adult age-progression," *Proc. Int'l Conf. Automatic Face and Gesture Recognition*, pp. 341–345, 2006.
[12] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convlutions," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 1–9, June 2015.

[13] S. Zagoruyko and N. Komodakis, "Wide residual networks," *CoRR*, vol. abs/1605.07146, 2016.
[14] A. Krizhevsky, I. Sutskever, and G.E. Hinton, "Imagenet classification with deep convolutional neural networks," *Proc. Annual Conf. Neural Information Processing Systems*, pp. 1–9, 2012.
[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 770–778, June 2016.
[16] M. Mathias, R. Beneson, M. Pedersoil, and L.V. Gool, "Face ditection without bells and whistles," *Proc. European Conf. Computer Vision*, vol. IV, pp. 720–735, 2014.