SN Computer Science (2020) 1:284
https://doi.org/10.1007/s42979-020-00294-w

ORIGINAL RESEARCH

Joint Age Estimation and Gender Classification of Asian Faces Using Wide ResNet

Hieu Trung Huynh[1,2] · **Hoang Nguyen**[1]

Received: 20 April 2020 / Accepted: 7 August 2020 / Published online: 27 August 2020

Abstract
Two key facial features, age and gender, have been widely explored. Companies and organizations have investigated in related applications in several fields including insurance, retails, marketing, etc. It would bring tremendous benefit, which allow companies to easily identify their customer demographics. Several approaches have been proposed with remarkable results. However, because of the lack of open and multi-ethnic datasets, most modern age and gender estimating models were trained solely based on white people with Western facial features, and thus fall short with non-Caucasian people. In this paper, we developed an applicable Wide ResNet model to estimate the age and the gender of Asian faces. The model was trained with a newly improved Asian face database. The experiments have shown promising results, as it can match the performance of Microsoft's how-old API estimator in a specific dataset.

Keywords Deep learning · Convolutional neural network · ResNet · Wide ResNet · Age and gender estimation

Introduction

The age and gender estimation problem has been investigated by several researchers due to its application ability in many fields such as insurance, retails, marketing, etc. Several industrial corporations are always actively searching for ways to utilize technologies to get an insight of their customer segments. Any company that comes into possession of such tools can harvest a massive amount of data, smoothly increase their revenues and benefits, surpass the competitors, and dominate the corresponding market. Another application of age and gender estimation is in surveillance systems, it can help to authorize people for buying restricted goods or adult products such as alcohol or tobacco. It can integrate with other biometric information to improve the accuracy of recognition systems. In addition, a recent report [1] has shown that the COVID-19 virus spread has the relevance of age, gender, and place of residence in population. Hence, these factors will provide the key information to guide the preparedness and response in healthcare facilities and the definition of policy interventions.

Collecting personal information on a large scale is not always easy, because it is relatively sensitive and not all people have the spare time to fill out multiple surveys at once. However, the task of facial image collection probably could be done simpler. It could be performed by installing several surveillance cameras around the facilities and it should be rather easy for companies and organizations to extract facial pictures of their visitors for security and research purpose. This may result in urgent needs for companies and organizations to have trust-worthy tools that can reasonably predict their visitor identity features, using just the faces.

In recent years, with the rise of deep learning and computer vision, researchers have been looking deeply into the age and gender estimation problem due to its practical influences. In this paper, we investigate in collecting and improving upon a multi-purpose Asian face dataset with the intention of partially addressing the problem of age and gender estimation of Asian people. Ultimately, the final aim of my paper is to make the best use of the data, as well as

This article is part of the topical collection "Future Data and Security Engineering 2019" guest edited by Tran Khanh Dang.

✉ Hieu Trung Huynh
hthieu@ieee.org

Hoang Nguyen
cs2014_hoang.ng@student.vgu.edu.vn

[1] Vietnamese-German University, Binh Duong new city, Binh Duong, Vietnam

[2] Industrial University of Ho Chi Minh City, Ho Chi Minh City, Vietnam

the existing deep learning techniques to put together a well-rounded implementation of Wide ResNet to create a reliable program that can be able to extract the age and gender of a certain Asian person with reasonable accuracy on both aspects.

The rest of this paper is organized as follows. Section 2 reviews related works for age and gender estimation. In Sect. 3, we propose a new approach for joint age estimation and gender classification. Experiments are described in Sect. 4. Results and discussion are presented in Sect. 5. Finally, we make a conclusion in Sect. 6.

Related Works

The problem of age and gender estimation is not new. It has been investigated by several teams, from which several different architectures have been proposed, including state-of-the-art ones. In [2], Chang-Ling Ku et al. proposed a method to examine many pictorial frames of a same person to make a more robust and stable prediction. Whereas in [3], Hayashi suggested that instead of completely leaving the feature-extraction step to the convolutional neural network (CNN) model as a black-box, it is advisable to manually extract facial identification landmarks such as wrinkles and use that to enhance the model performance, as they are indications of age. Those are all plausible approaches to such problem. Tian and Chen [4] proposed a joint learning method for age and gender estimation based on support vector machine (SVM) and ordinal regression methods. Four datasets were evaluated including FG-NET, morph album I, album II, and images of groups. Another approach based on support vector machine was proposed by Aswathy et al. [5]. This approach estimates the facial attributes using drop-SVM and kNN from the images of faces acquired in the wild conditions.

In [6] and [7], Rothe et al. came up with a revolutionary method called Deep EXpectation (DEX). The method is strikingly simple. In the beginning stage, the faces were cropped and aligned using available toolkits, and then, a regular VGG-16 network was used for the age detection as a classification task with 101 classes (representing the age from 0 to 100). Finally, after receiving the Softmax distribution for the 101-dimension vector, the output values were determined by performing a dot product operation between that vector and another vector containing discrete values from 0 to 100. It is argued that this process is more robust and stable than both the logistic regression and the linear regression approach, while yielding a better result at the same time. As presented in [7], authors claimed that their implementation is the current state-of-the-art in predicting the subject's apparent age, as tested on the IMDB-WIKI dataset [6, 7]. These two papers are the main academic resources whose concept we will utilize into the model to achieve the desirable results. Smith and Chen [8] proposed a method using transfer learning with deep CNNs. The transfer learning is based on VGG19 and VGGFaces with deep CNN hierarchy. First, the subject is classified based on gender. The male and female models were used to predict the age. This approach was evaluated on MORPH II dataset [9]. Most modern Age and Gender estimating models were trained based on white people with Western facial features, and the application on Asian faces has limitations.

Proposed Method

Our proposed scheme is depicted in Fig. 1. The training process is described in Fig. 1a, which includes two stages. The first stage is the image augmentation by using Random erasing and Mixup processes. The Wide ResNet model is trained in the second stage. The trained network is then used for age estimation and gender classification, as shown in Fig. 1b.

Regularization and Image Augmentation

To prevent the model from overfitting, conventional regularization and image augmentation should be applied, and they may consist of the L_1, L_2 regularization, batch normalization, dropout, image shifting, rotating, and flipping. However, from experiments, it came to our attention that the model still heavily suffers from the state of overfitting. We eventually tackle the troublesome problem using random erasing and mixup.

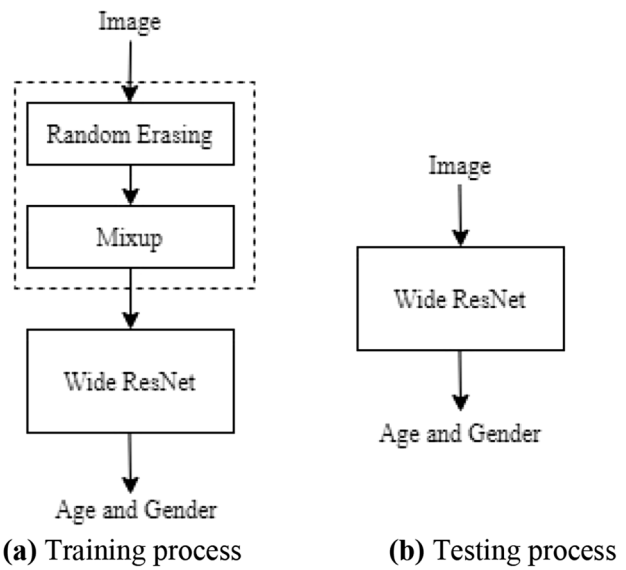


Fig. 1 The proposed scheme for age and gender estimation. The left is training process and the right is testing process. **a** Training process. **b** Testing process

Random erasing: The idea behind this method is truly intuitive. It would observe that human can still perceive the information conveyed by the image even if there are slight distortions, as long as the distorted portion is not too big that it covers important structure parts. A powerful model should be able to perform at the same level, grasp the context of an image from structure of the overall object architecture, rather than each individual pixel in block. The random erasing was introduced to assist model in reaching that intelligent stage. In which, any image will either be kept unchanged or randomly have a rectangle region of an arbitrary size assigned with arbitrary pixel values. By simply introducing a reasonable noise to the training sample, this method may enhance the model to become less prone to overfitting. We enforced the algorithm with similar set of hyper-parameters as concluded in [10]. To further elaborate, the process can be described as a sequence of mathematical operations. Before training process, each image I has a probability of p of activating random eraser, and inherently, a probability of $1 - p$ to remain unchanged. The original size of the image is:

$$S = W \times H, \tag{1}$$

where W and H are the width and height of image, respectively. A randomly selected rectangle region I_e with area ratio s_a is chosen between s_l and s_h. An aspect ratio r_e is also uniformly randomly initialized between r_1 and r_2. The area of I_e can be computed as:

$$S_e = s_a \times S, \tag{2}$$

which is then used to calculate:

$$H_e = \sqrt{S_e \times r_e} \tag{3}$$

$$W_e = \sqrt{\frac{S_e}{r_e}}. \tag{4}$$

Afterwards, another point $P = (x_e, y_e)$ is arbitrarily chosen within I, and the region $I_e = (x_e, y_e, x_e + W_e, y_e + H_e)$ is then set to a certain value in the range [0, 255]. If I_e falls outside I, P is re-selected until the requirement is satisfied. In our implementation, we set $p = 0.5$, $s_l = 0.02$, $s_h = 0.4$, $r_1 = 0.3$, $r_2 = 1/0.3$, similar to [10]. The resulting images are forwarded to the next stage. An example of random erasing is shown in Fig. 2. In Fig. 2b, a random portion is erased, but it affects very little our perception of the subject's age and gender.

After pre-processing with random erasing, an extra measurement is taken to further guarantee the generalization of the same model. In essence, rather than handling each image independently, Mixup [11] groups pairs of samples and their labels together before training the model on a convex combination of such pairs. A weakly trained model will design an overly complex mapping between an image and its label, and then fail to recognize very similar images with slight modification or adversarial generation. To tackle such problem, the Mixup helps regularizing the neural network to favor simple linear behavior in-between training sample instead of complex correlations. The method is proven to enhance robustness, stability, and accuracy of models on popular dataset like ImageNet, CIFAR-10, and CIFAF-100 dataset, resulting in state-of-the-art architecture. Despite its powerful capability, Mixup can be easily implemented within a few lines of code with minimal overhead. To put it briefly, Mixup creates virtual training samples through the formula:

(a) Original image **(b)** Random erasing image

Fig. 2 Example of random erasing on face dataset. Here, the random erased portion affects very little on our perception of the subject's age and gender, which is 26, male. **a** Original image. **b** Random erasing image

$$\tilde{\mathbf{x}} = \lambda \mathbf{x}_i + (1 - \lambda)\mathbf{x}_j \tag{5}$$

$$\tilde{y} = \lambda y_i + (1 - \lambda) y_j, \tag{6}$$

where $\mathbf{x}_i$ and $\mathbf{x}_j$ are raw input vectors; y_i and y_j are one-hot encoding labels; $(\mathbf{x}_i, y_i)$ and $(\mathbf{x}_j, y_j)$ are two randomly drawn examples from the mini training batch, and $\lambda \in [0, 1]$. In the proposed model, for simplicity purposes, we draw λ from a Beta distribution with $\alpha = \beta = 0.2$, $\lambda \sim \text{Beta}(0.2, 0.2)$ [11]. The output is now ready to be fed into the model for training. An example of random erasing and Mixup is shown in Fig. 3.

Wide ResNet Architecture

The residual networks (ResNet) have achieved state-of-the-art in several benchmarks, including object classification and detection [12, 13]. The order of activations in the residual networks representing identity mappings in the residual blocks may affect the training performance. The residual block with identity mapping is represented mathematically as in Eq. (7) [13]:

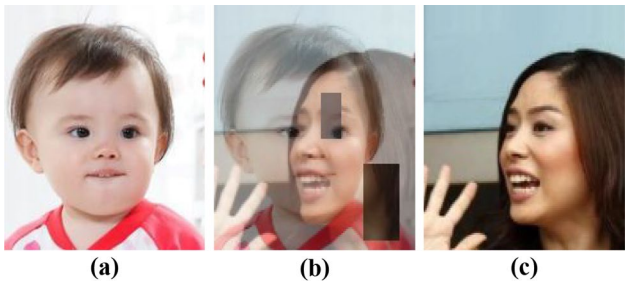


Fig. 3 Example of random erasing + Mixup. After being randomly erased, the two images on the left (**a**) and right (**c**) are mixed up with $\lambda = 0.5$ to create the middle image (**b**). Note that their one-hot labels are also combined

$$\mathbf{o}_{l+1} = \mathbf{o}_l + F(\mathbf{o}_l, W_l), \tag{7}$$

where $\mathbf{o}_l$ and $\mathbf{o}_{l+1}$ are the input and output of the lth block in the network, and F and W_l are residual function and its parameters. Various residual blocks can be used including a combination of convolutional layers with batch normalization (BN) and ReLU preceding convolution. To increase the representational power of residual block, the popular methods could be applied including (1) add more convolutional layers per block; (2) increase the filter sizes in convolutional layers; or (3) widen the convolutional layers by extending feature channels.

Regarding our model architecture for age and gender estimation, we have utilized the variation of the popular ImageNet-champion residual network, called wide residual network, or wide ResNet. It was introduced by Zagoruyko and Komodakis in [14]. The fundamental concept lies behind this network is that in addition to making the use of the skip connection like the regular ResNet, the order of layers in a block is rearranged from conv-BN-ReLU to BN-ReLU-conv to enhance its feature-extraction capability. It is also expanded in terms of the feature channels by a widening factor k. As for the rearrangement of the layer order, Authors [14] showed through experimental results that the new order, indeed, executes faster than the original one while achieving a better accuracy level.

In terms of the feature channel extension, there is an ongoing debate on whether the width or the depth of the network contributes more to the overall success of the model. Conventionally, the original ResNet model suggests that to achieve higher accuracy, ones could just simply increase the deepening factor N, also referred to as the number of blocks in a certain stage, or the number of stages in general without having worry about gradient vanishing/exploding like older architecture due to the powerful skip connection. In other words, keeping on stacking more layers until reaching a desirable result was the most straightforward method. However, deeper network also subsequently increases the number of parameters in linear time and may make it longer and harder to converge. Furthermore, there is also the problem of diminishing feature reuse (a few residual blocks learning useful representations or many blocks sharing very little information with small contribution to the final goal) in very deep and thin residual networks.

Table 1 Structure of wide residual networks

Group name	Output size	Block type = B(3,3)
Conv1	[64×64, 16]	[3×3, 16]
Conv2	[64×64, 16×k]	$\begin{bmatrix} 3\times3\ 16\times k \\ 3\times3\ 16\times k \end{bmatrix}\times N$
Conv3	[32×32, 32×k]	$\begin{bmatrix} 3\times3\ 32\times k \\ 3\times3\ 32\times k \end{bmatrix}\times N$
Conv4	[16×16, 64×k]	$\begin{bmatrix} 3\times3\ 64\times k \\ 3\times3\ 64\times k \end{bmatrix}\times N$
Avg-pool	[16×16, 64×k]	[8×8]
Flatten	[1×16×16×64×k]	N/A
Fc-gender	[1×2]	N/A
Fc-age	[1×101]	N/A

The Network width is determined by factor k, here $k = 8$. Groups of convolutions are shown in brackets where N is the number of blocks in group, here $N = 2$. Downsampling is performed by the first layers in groups conv3 and conv4. The final two fully connected layers perform prediction based on the flatten layer

Thus, instead of senselessly stacking more layers and waste computational resources, a Wide ResNet suggests addressing the above problems by increasing the number of channels by a factor of k. It allows each block to learn a more meaningful feature representation of the data. Although this makes the total number of parameters grows in quadratic time, most modern GPUs are programmed to make parallel computation on large tensors for efficiencies, hence reducing the total amount of time needed for training a model [15]. Furthermore, adding a simple dropout layer between two BN-ReLU-conv sequences also gives more room for model generalization. The specific architecture is described in Table 1.

Similar to the original Wide ResNet, our implementation begins with a 2D convolutional layer to extract the early features, with a 3×3 convolution block and a 16 feature channels. The main difference with our model is that instead of taking the input shape 32×32×3, we decided to double it to 64×64×3, since the Megaage_Asian dataset [16] contains pictures with higher resolution, which gives clearer edge and facial feature to detect. Following that, there are three main groups; each group consists of two convolutional blocks (here, $N = 2$) and an additional dropout layer with the dropout probability of 0.2 between those blocks to prevent overfitting. Each block contains one BN layer, one ReLU layer, and one convolutional layer. During the transition between

two groups, the size of features is halved, while the channel planes are doubled. In the case where the input and output size do not share the same dimension, a 1D convolutional layer is applied to resize them for matching dimensionality. Finally, after all the convolutional groups have been performed, the data then flow through one average pooling layer (with block size of 8×8) and flattened out. Two final fully connected layers are then applied simultaneously to the flatten feature vector to perform both the age estimation and gender classification. The Gender classification layer is a 2-dimensional Softmax vector, representing the distribution probability that the person is either a male or a female. Similarly, the Age estimation vector is 10-dimensional, relative to the possible human age between 0 and 100 (here we presumed that people only live up to 100 year for simplicity).

Experiments

Dataset Description

Conventionally, most research teams refer to the IMDB—WIKI dataset when they tackle the age and gender detection problem, since it is among the most popular public datasets, with 523,051 images of celebrities along with their age and gender. However, it only contains an extremely small portion of people with Asian background. Therefore, models trained on that dataset unsurprisingly underperformed when presented with Asian descendants. Considering that Asian people makes up to ~ 60% of the world population, that is indeed a serious downfall for such models. Consequently, we did a handful of research and finally came across the Megaage_Asian dataset collected by MMLAB, the Chinese University of Hong Kong or SenseTime [16]. The dataset consists of 40,000 up-close, frontal, and non-blurry Asian facial images, along with the subject's age and gender, stored in a separate.txt file. The images in the dataset mostly belong to people from different Asian ethnicity such as Thailand, Vietnam, or Japan, cropped and aligned centering around their faces, varying from different angles and light direction. The age of the subject ranges between 1 and 70 which is truly diversified. Some samples from the dataset are shown in Fig. 4. In the figure, the first string shows the gender of subject that is either Male or Female, and the following number is the subject's age. This format follows for subsequent figures as well if not otherwise specified.

For clearance, identifying subject's gender is not as challenging as figuring out subject's age, as a normal human can execute the former task almost flawlessly. Although our gender labels are not professionally verified, it still holds certain merits. The dataset consists of 40 thousand images for training, and 3945 other images for testing, which is a reasonable amount for the task at hands. The female:male ratio is 20,684:23,261, or roughly 9:10. The distribution of the age range can be interpreted from Fig. 5.

Implementation Details

The entire Wide ResNet is implemented using Python programming language along with its libraries such as OpenCV, NumPy, and Keras deep learning framework, trained on Amazon Web Service (AWS), with 8 GB of GPU and 60 GB

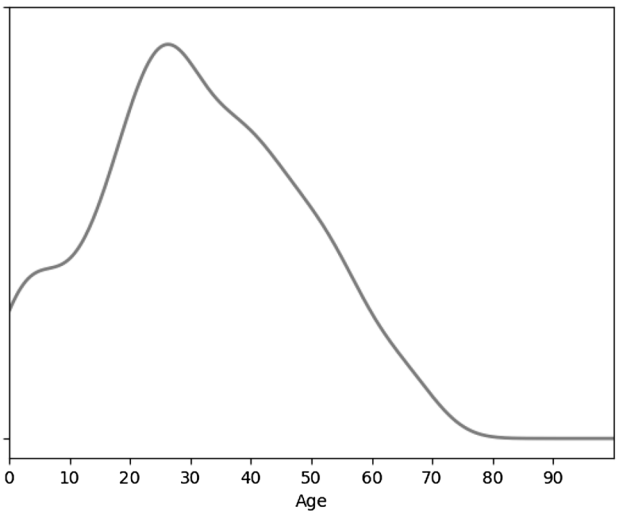


Fig. 5 Age distribution of the dataset. Although it is not exactly Uniform, the dataset still offers people with a variety of age range

(a) Female, 1 **(b)** Male, 14 **(c)** Female, 41 **(d)** Male, 67

Fig. 4 Samples from the Megaage_Asian dataset. The dataset is truly diversified, as it contains people with different ages, genders, and ethnicities. **a** Female, 1. **b** Male, 14. **c** Female, 41. **d** Male, 67

of Memory. The entire code and our implementations can be found at: https://github.com/amidadragon/asian-age-gender-estimation-2.

The complete workflow of our methods is described as following. To start, all images are rescaled to the size of 64×64. Afterwards, the dataset is separated into training set and validation set with the ratio of 9:1, which has 36,000 and 4000 images relatively. Each mini-batch of 32 training samples consecutively undergoes a pre-processing series of random erasing and Mixup before feeding onto the network. Regarding the Wide ResNet, the parameters are randomly initialized using the He random formula [17]:

$$w^{[l]} = X \times \sqrt{\frac{2}{n^{l-1}}}, \tag{8}$$

where $w^{[l]}$ is the parameters of layer l, $X \sim N(0,1)$ and n^{l-1} is the number of neurons in the previous layer (layer $l-1$). This initialization method allows models to converge to a global minimum faster and more efficient. The weights are optimized by the mini-batch gradient descent with the initialized learning rate α set as 0.1; it is decayed overtime through the formula:

$$\alpha_{\text{new}} = \frac{\alpha_{\text{prev}}}{1 + \text{decay\_rate} \times \text{epoch\_nums}}, \tag{9}$$

where α_{new} is new learning rate, α_{prev} is the previous learning rate, epoch_nums is the epoch number, and decay_rate is the decay rate and set as $1e^{-6}$. After each epoch, the model is freshly tested on a new validation set that it has never observed before to evaluate its generalization power and make necessary adjustments. Only the model which has the best scores on the validation dataset is saved and used for further inspections.

Results and Discussion

The training history of the Wide ResNet model is depicted in Fig. 6. It can be interpreted from the figure that the gender validation accuracy peaked at 92%, which is a rather reasonable result compared to the other proposed methods. However, the fascinating discovery lies within the age classification. Its accuracy peaks at 15% around epoch 80, but that is not the metrics that we intent to examine for the age estimation problem, since precisely predicting a person age is a challenging task for any person, or even professional anthropologists. Hence, we utilize the DEX method mentioned in [6, 7] to calculate the predicted age value. In terms of the DEX method, the Softmax distribution only serves as an intermediate means to determine the final age prediction, unlike the traditional estimation models.

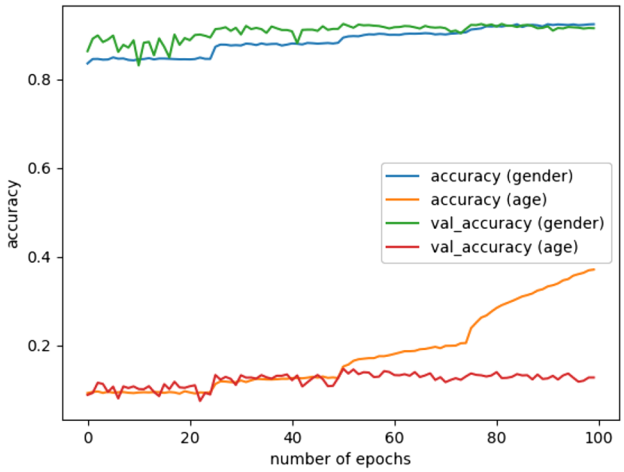


Fig. 6 Training process of Wide ResNet. Keep in mind that the validation accuracy (the green and red line) are the metrics that ultimately matters

The expected value for the subject's age is calculated by the formula:

$$E(\mathbf{o}) = \sum_{i=0}^{100} y_i o_i, \tag{10}$$

where $\mathbf{o} = \{o_0, o_1, \ldots, o_{100}\}$ is the Softmax output probabilities and y_i's are the discrete years, ranging from 0 to 100, corresponding to each class i. $E(\mathbf{o})$ is the final prediction value for a subject's age. With $E(\mathbf{o})$ at our disposal, we compute the MAE over the 4000 testing images by:

$$\text{MAE} = \frac{\sum_{i=1}^{n} \left| \hat{y}_i - E_i(\mathbf{o}) \right|}{n}, \tag{11}$$

where n is the number of images, and $\hat{y}_i$ and $E_i(\mathbf{o})$ are the actual and predicted age of person in the image i, respectively. For the validation set, MAE $\approx$ 4.2. To avoid bias in the model selection step, we also perform the exact measurement with the 3945 testing images, which yields MAE = 4.05. The result indicates that, on average, the model misses a subject's age by a margin of roughly 4 years. It may note that, excluding the young childhood phase when the face might alter dramatically, generally person face does not change too much in a 4-year period.

It may state that the gender aspect is easier to predict, because there exist only two options to choose, either male or female, and it is also an effortless task for human to perform with very high accuracy. However, the age feature is definitely significantly harder to recognize, because there are more than 100 possible cases. Each individual person has different facial features, depending on their living

(a) 1, 1 **(b)** 13, 11 **(c)** 30, 28 **(d)** 51, 53

Fig. 7 Some accurately predicted images. The number on the left is the subject actual age, and the right is that of our prediction model. **a** 1, 1. **b** 13, 11. **c** 30, 28. **d** 51, 53

(a) 12, 21 **(b)** 17, 9 **(c)** 20, 13 **(d)** 33, 40

Fig. 8 Inaccurately predicted images. **a** 12, 21. **b** 17, 9. **c** 20, 13. **d** 33, 40

Table 2 Age groups

Group 0	Group 1	Group 2	Group 3	Group 4	Group 5
0–11	12–18	19–25	26–32	33–39	40–100

conditions and ethnicity. In addition, the light direction and facial angle could also affect how old a certain people may look.

Some images where our model makes the most plausible guesses are shown in Fig. 7, and the predicted and actual ages are given in the first and second number, respectively. It can be observed that they have natural filter and light, as well as neutral emotion.

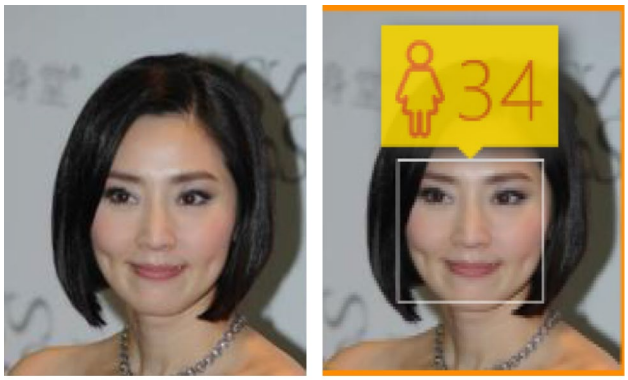
There are still some images which result in low accuracy by our model, as shown in Fig. 8, and the predicted and actual ages are given in left and right values, respectively. Some faces emerging such images make up unnaturally or have a big smile. They result in more wrinkles and eventually give out a feeling that individuals in the image appear different from they actually are. It could also be that they just possess different facial features in general. On the other hand, there are also images with heavy photoshop, resulting in noticeably and undeniably younger.

Since an MAE score can be difficult to fully interpret, we also trained the same model, with similar hyper-parameters on the same data with age mapped to a certain age range, based on different phases of a human life circle, as shown in Table 2. It includes toddlers, teenagers, young adults, adults, midlife, and seniors.

The new model achieved an accuracy level of 67.66% and a 96.3% 1-off accuracy, and the confusion matrix is illustrated in Fig. 9. From the confusion matrix, we can infer that most of the mistakes are between group 2 (19–25) and group 3 (26–32). A reasonable explanation would be that, at these phases, they reach a certain maturity and stable facial

Fig. 9 Confusion matrix

True group/ Predicted group	0	1	2	3	4	5
0	**1133**	95	18	5	2	0
1	69	**286**	158	15	3	0
2	6	79	**431**	219	15	3
3	1	8	159	**368**	95	14
4	2	1	31	129	**181**	61
5	0	1	2	23	66	**270**



(a) Female, 33 **(b)** Female, 34

Fig. 10 Our model (left) vs Microsoft's model (right). My model matched that of Microsoft in some instances. She subject is actually 38 years old. **a** Female, 33. **b** Female, 34

feature, and do not change as drastically as other phases, and hence, the confusion of the model is similar to human. Toddlers and teenagers, on the other hands, go through dramatic changes with their hormones level, resulting in distinct facial features.

With our best knowledge, it is difficult to find the previous studies which have run on experiment with the same testing dataset. Authors in [6] and [7] provide that they achieved the MAE score of 5.0 on the IMDB-WIKI dataset, but that is not to be compared to ours, since their data are much larger, and also, the distribution is differ greatly. However, Microsoft published a similar API to predict age and gender of a certain person through just one submitted image. We requested the API to check out its accuracy on the same testing dataset. The result is quite surprising as it turned out. Our model got 1% better than the Microsoft in terms of gender accuracy, and 0.2 less of in term of MAE score. This is not in anyway a statement that our model outperforms those of Microsoft. Microsoft Deep Leaning model presumably generalizes better to non-Asian individuals, while our model would not make such rivalry prediction. However, with a limited amount of only 40,000 images, and 48 h of GPU training, we were able to partially match the performance of Microsoft, who has in their possession a big dataset of images and GPU computational power. This could place a foundation for future development using similar network architecture (Fig. 10).

Conclusion

The age and gender estimation could be applied in many real-world applications, it has been investigated by several researchers and several models have been proposed. However, they have limitations relating to population. In this study, we have introduced an approach to partially addressing that problem, by focusing only on the Asian population. The proposed approach utilizes the state-of-the-art deep learning and computer vision algorithms to achieve high accuracy. The well-rounded Wide ResNet model and image augmentation techniques have been applied. Experimental results shown that the proposed approach obtains promising outcomes with a small error rate and a reasonable accuracy level. The model could easily be applied for non-Asian population.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Goujon A, Natale F, Ghio D, Conte A, Dijkstra L. Age, gender, and territory of COVID-19 infections and fatalities. EUR 30237 EN, Publications Office of the European Union, 2020.
2. Ku C-L, Chiou C-H, Gao Z-Y, Sai T-J, Fuh C-S. Age and gender estimation using multiple-image features. Biometr Recogn. 2013;441–448.
3. Hayashi J-I, Koshimizu H, Hata S. Age and gender estimation based on facial image analysis. Knowl Based Intell Inform Eng Syst. 2003;863–869.
4. Tian Q, Chen S. Joint gender classification and age estimation by nearly orthogonalizing their semantic spaces. Image Vis Comput. 2018;69:9–21.
5. Unnikrishnan A, Ajesh F, Kizhakkethottam JJ. Texture-based estimation of age and gender from wild conditions. Procedia Technol. 2016;24:1349–57.
6. Rothe R, Timofte R, Gool LV. DEX: Deep EXpectation of apparent age from a single image. IEEE Intern Conf Comput Vis-Worksh (ICCVW). 2015;12:1–8.
7. Rothe R, Timofte R, Gool LV. Deep expectation of real and apparent age from a single image without facial landmarks. Intern J Comput Vis (IJCV). 2018;126:144–57.
8. Smith P, Chen C. Transfer learning with deep CNNs for gender recognition and age estimation. In: The 5th National Symposium for NSF REU Research in Data Science, Systems, and Security, pp 1349–1357, 2018.
9. Ricanek K, Tesafaye T. MORPH: a longitudinal image database of normal adult age-progression. In: 7th International Conference on Automatic Face and Gesture Recognition (FGR06) 2006.
10. Zhong Z, Zheng L, Kang G, Li S, Yang Y. Random erasing data augmentation. CoRR. vol. abs/1708.04896, 2017.
11. Zhang H, Cissé M, Dauphin YN, Lopez-Paz D. mixup: beyond empirical risk minimization. CoRR. vol. abs/1710.09412, 2017.
12. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2016.
13. He K, Zhang X, Ren S, Sun J. Identity mappings in deep residual networks. In: Computer Vision-ECCV 2016. ECCV 2016. Lecture Notes in Computer Science 2016.
14. Zagoruyko S, Komodakis N. Wide residual networks. CoRR. vol. abs/1605.07146, 2016.

15. Dietz M. Understand deep residual networks—a simple, modular learning framework that has redefined state-of-the-art, Medium, 2017.
16. Zhang Y, Liu L, Li C, Loy C.C. Quantifying facial age by posterior of age comparisons, 2017.
17. He K, Zhang X, Ren S, Sun J. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. CoRR. vol. abs/1502.01852, 2015.