AGE GROUP CLASSIFICATION IN THE WILD WITH DEEP ROR ARCHITECTURE

A. Ke Zhang, B. Liru Guo, C. Ce Gao D. Zhenbing Zhao*

North China Electric Power University
Department of Electronic and Communication Engineering
Baoding, Hebei, China

E. Miao Sun, F. Xingfang Yuan

University of Missouri
Department of Electrical and
Computer Engineering
Columbia, MO, USA

ABSTRACT

Automatically predicting age group from face images acquired in unconstrained conditions is an important and challenging task in many real-world applications. Nevertheless, the conventional methods with manually-designed features on in-the-wild benchmarks are unsatisfactory because of incompetency to tackle large variations in unconstrained images. In this paper, we propose a new CNN based method for age group classification leveraging Residual Networks of Residual Networks (RoR), which exhibits better optimization ability for age group classification than other CNN architectures. Moreover, two modest mechanisms based on observation of the characteristics of age group are presented to further improve the performance of age estimation. Our experiments illustrate the effectiveness of RoR method for age estimation in the wild, where it achieves better performance than other CNN methods. Finally, the Pre-RoR-58+SD with two mechanisms achieves new state-of-the-art results on Adience benchmark.

Index Terms— Age group classification, RoR, Pretraining, Weighted loss

1. INTRODUCTION

Age, one of the key facial attributes, plays very foundational roles in social interactions, making age estimation from a single face image an important task in intelligent applications, such as access control, human-computer interaction, law enforcement, marketing intelligence and visual surveillance, etc [1].

Over the last decade, most methods used manually-designed features to estimate age [2, 3, 4, 5, 6, 7], and they achieved respectable results on the benchmarks of *constrained* images, such as FG-NET [8] and MORPH [9]. However, manually-designed features based methods behave unsatisfactorily on recent benchmarks of *unconstrained* images, namely “in-the-wild” benchmarks, including Public Figures [10], Gallagher group photos [11] and Adience [12] for these features’ ineptitude to approach large variations in appearance, noise, pose and lighting.

Deep learning, especially deep Convolutional Neural Networks (CNN) [13, 14, 15, 16, 17], has proven itself to be a strong competitor to the more sophisticated and highly tuned methods. Although unconstrained photographic conditions bring about various challenges to age prediction in the wild, we can still enjoy great improvements brought by CNNs [18, 19, 20, 21, 1]. The optimization ability of neural networks is critical to the performance of age estimation, while existing CNNs designed for age estimation only have several layers, which severely limits the development of age estimation. Therefore, we construct a very deep CNN, Residual networks of Residual networks (RoR) [22], for age group estimation in the wild. To begin with, we construct RoR with basic residual block for age group classification. In addition, analysis of the characteristics of age estimation suggests two modest mechanisms, pre-trained CNN by gender and weighted loss layer, to further increase the accuracy of age estimation. Finally, through massive experiments on Adience dataset, our RoR model achieves the new state-of-the-art results on Adience dataset.

2. RELATED WORK

In the past twenty years, human age estimation from face image has benefited tremendously from the evolutionary development in facial analysis. Early methods for age estimation were based on geometric features calculating ratios between different measurements of facial features [23]. However, these pixel-based methods are not suitable for in-the-wild images which have large variations in pose, illumination, expression, aging, cosmetics and occlusion. After 2007, most existing methods used manually-designed features in this field, such as LBP [2], SFP [3], and BIF [4]. Based on these manually-designed features, regression and classi-

*Thanks to National Natural Science Foundation of China (Grants No.61302163, No.61302105, No.61401154 and No.61501185), Hebei Province Natural Science Foundation (Grants No.F2015502062, No. F2016502101 and No.F2016502062) and the Fundamental Research Funds for the Central Universities (Grants No.2016MS99 and No.2015ZD20) for funding.

978-1-5090-2175-8/17/$31.00 ©2017 IEEE

fication methods are used to predict the age of face images. SVM based methods [4, 12] are used for age group classification. For Regression, linear regression [5], SVR [6], and CCA [7] are the most popular methods for accurate age prediction. However, all of these methods were only proven effective on constrained benchmarks, and could not achieve respectable results on the benchmarks in the wild.

Recent research on CNN showed that CNN model can learn a compact and discriminative feature representation when the size of training data is sufficiently large, so an increasing number of researchers start to use CNN for age estimation. Yi et al. [18] first proposed a CNN based age estimation method, Multi-Scale CNN. Wang et al. [19] extracted CNN features, and employed different regression and classification methods for age estimation on FG-NET and MORPH. Levi et al. [20] used CNN for age classification on unconstrained Adience benchmark. With the development of deeper CNNs, Hou et al. [21] proposed a VGG-16-like model with Smooth Adaptive Activation Functions (SAAF) to predict age group on Adience benchmark. VGG-16 architecture and SVR were used for age estimation on top of the CNN features. Deep EXpectation (DEX) formulation [1] was proposed for age estimation based on VGG-16 architecture and a classification followed by a expected value formulation. DEX achieves state-of-the-art results on several constrained or unconstrained standard benchmarks.

3. METHODOLGY

3.1. Network architecture

RoR [22] is based on a hypothesis: The residual mapping of residual mapping is easier to optimize than original residual mapping. To enhance the optimization ability of residual networks, RoR can optimize the residual mapping of residual mapping by adding shortcuts level by level based on residual networks.

In order to train the high-resolution Adience dataset, we first construct RoR based on the Pre-ResNets for Adience, and denote this kind of RoR as Pre-RoR. Pre-ResNets [24] include two types of residual block designs: basic residual block and bottleneck residual block. We choose basic residual block in this paper. Fig. 1 shows the Pre-RoR constructed based on original Pre-ResNets with L basic blocks. The shortcuts in these L original residual blocks are denoted as the final-level shortcuts. To start with, we add a shortcut above all basic blocks, and this shortcut can be called root shortcut or first-level shortcut. We use 64, 128, 256 and 512 filters sequentially in the convolutional layers, and each kind of filter has different number (L_1, L_2, L_3, L_4, respectively) of basic blocks which form four basic block groups. Furthermore, we add a shortcut above each basic block group, and these four shortcuts are called second-level shortcuts. Then we can continue adding shortcuts as the inner-level shortcuts. Lastly, the shortcuts in basic residual blocks are regarded as the final-level shortcuts. Let m denote a shortcut level number. In this paper, we choose level number m=3 according to the analysis of Zhang et al. [22], so the RoR has root-level, middle-level and final-level shortcuts, shown in Fig. 1.

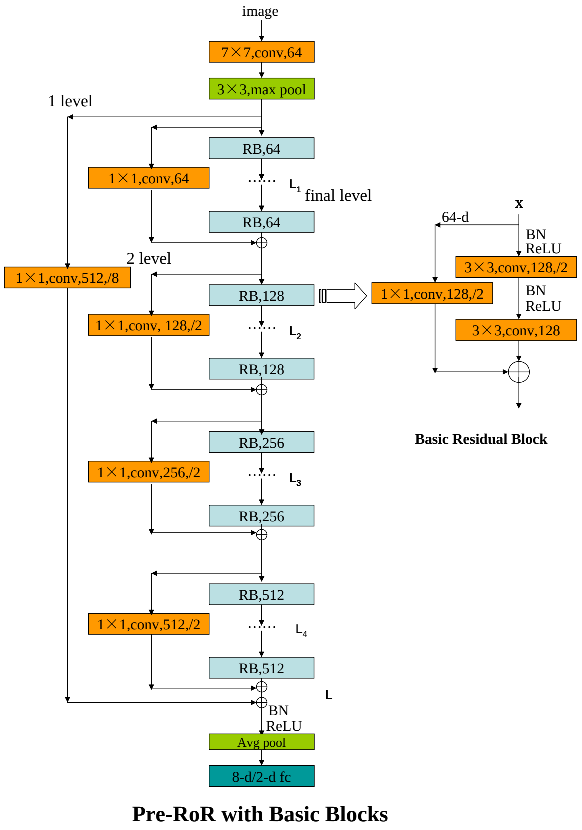

Fig. 1. Pre-RoR architecture.

3.2. Pre-training with gender

Like face recognition, age estimation can be easily affected by many intrinsic and extrinsic factors. Some of the most important factors include identity, gender and ethnicity, together with other factors like pose, illumination and expression (PIE). We can alleviate the effects of these factors by using large datasets in the wild, but the existing datasets for age estimation are generally relatively small. To some extent, gender affects age judgments. On the one hand, the aging process of men slightly differs from women due to different longevity, hormones, skin thickness, etc. On the other hand, women are more likely to hide their real age by using makeup. So real-world age estimations for men and women are not exactly the same. Guo et al. [7] first manually separated the dataset according to the gender labels, then trained an age estimator on each subset separately. Inspired by this, we train CNN by gender initially, then replace the gender prediction layer with age prediction layer, and fine-tune the whole CNN structure at last.

3.3. Training with weighted loss layer

There are some diversities lying between general image classification and age estimation. Firstly, the different classes in general image classification are uncorrelated, but the age groups have a sequential relationship between labels. These interrelated age groups are more difficult to distinguish. Secondly, human aging processes show variations in different age ranges. For example, aging processes between mid-life adults and children are not equivalent. In this paper, we will analyze the law of human aging, and do age estimation under its guidance. For human, it is easier to distinguish who is the older one out of two people than to determine the persons' actual ages. Based on this characteristic and age-ordered groups, we define y_i, i=1,2...,K, where K is the number of age group labels. Then for a given age group $k \in K$, we separate the dataset into two subsets X_k^+ and X_k^- as follows:

$$\begin{aligned} X_k^+ &= \{(x_i,+1)|y_i > k\} \\ X_k^- &= \{(x_i,+1)|y_i \le k\} \end{aligned} \tag{1}$$

Next, we use the two subsets to learn a binary classifier that can be considered as a query: "Is the face older than age group k?" There are eight classes (0-2, 4-6, 8-13, 15-20, 25-32, 38-43, 48-53, 60-) in Adience dataset, so we can choose k=1,2,...,7. By doing so, we get seven binary-class datasets, and the results of these binary classifiers can form a human aging curve which represents the human aging process. We execute some experiments on folder0 of Adience dataset with 4c2f CNN described in [20] (just using two classes instead of eight classes), and the aging curve is described in Fig. 2 We discover that the 4th, 5th and 6th results are smaller than the others. As a conclusion, the aging process of smaller and greater age group is faster than intermediate age groups, so it is harder to distinguish intermediate age groups comparing to smaller and greater age groups. Through above analysis, we realize the 4th, 5th, 6th and 7th groups are more difficult to estimate, so we apply higher loss weights to these age groups. Based on several experiments, we choose (1,1,1,1.3,1.5,1.5,1.3,1) as the loss weight distribution for the eight age groups.

4. EXPERIMENTS

In this section, extensive experiments are conducted to present the effectiveness of the proposed RoR architecture and two mechanisms.

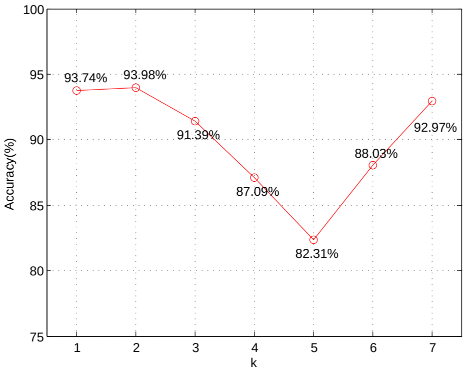

Fig. 2. The aging curve by binary classifiers.

4.1. Implementation

For Adience dataset [12], we do experiments by using 4c2f-CNN [20], VGG [15], Pre-ResNets [24] and our Pre-RoR architectures, respectively. We use the basic block Pre-ResNets in [24] to construct RoR architecture. Pre-RoR includes Pre-RoR-34 (34 layers), Pre-RoR-58 (58 layers) and Pre-RoR-82 (82 layers). Each residual block group in different Pre-RoR has different number of residual blocks, as shown in Table 1. In order to get the optimal model of Pre-RoR on Adience dataset, we adopt Stochastic Depth (SD) method [25] and Type A+B of shortcut type in Pre-RoR model according to the analysis of Zhang et al. [22], and denote this Pre-RoR as Pre-RoR+SD.

Number of Layers	Number of blocks in each Group
34	3, 4, 6, 3
58	5, 6, 12, 5
82	7, 8, 14, 7

Table 1. The number of residual blocks.

Our implementations are based on Torch 7 with one Nvidia Geforce Titan X. We initialize the weights as in [17]. We use SGD with a mini-batch size of 64 for these architectures. The total epoch number is 164. The learning rate starts from 0.1, and is divided by a factor of 10 after epoch 80 and 122. We use a weight decay of 1e-4, momentum of 0.9, and Nesterov momentum with 0 dampening [26]. For stochastic depth drop-path method, we set p_l with the linear decay rule of p_0= 1 and p_L=0.5.

The entire Adience collection includes 26,580 256×256 color facial images of 2,284 subjects, with eight classes of age groups and two classes of gender. Testing for age group classification is performed using a standard five-fold, subject-exclusive cross-validation protocol, defined in [12]. For data augmentation, VGG, PreResNets and Pre-RoR use scale and aspect ratio augmentation [26] instead of scale augmentation used in 4c2f-CNN.

4.2. Age group classification by Pre-RoR

We do several experiments of age group classification with standard five-fold, subject-exclusive cross-validation protocol. And we report the **exact accuracy**(correct age group predicted) and **1-off accuracy** (correct or adjacent age group predicted) as [12]. The age cross-validation results of Pre-RoR+SD with different depths are shown in Table 2. Generally, ResNets [17] and RoR [22] can improve performance by increasing depth. We estimate age by Pre-RoR-34+SD, Pre-RoR-58+SD and Pre-RoR-82+SD. The age estimation result of Pre-RoR-58+SD is better than Pre-RoR-34+SD, but Pre-RoR-82+SD is worse than Pre-RoR-58+SD, which is caused by degradation. Our Pre-ROR-58+SD achieves the best performance, which outperforms 4c2f-CNN by 18.8% and 5.7% on exact and 1-off accuracy of Adience dataset.

Method	Exact Acc(%)	1-off(%)
4c2f-CNN	52.62±4.37	88.61±2.27
VGG-16	54.64±4.76	54.64±4.76
Pre-ResNets-34	60.15±3.99	90.90±1.67
Pre-RoR-34+SD	62.35±4.69	93.55±1.90
Pre-RoR-58+SD	62.50±4.33	93.63±1.90
Pre-RoR-82+SD	62.14±4.10	93.68±1.22

Table 2. The age cross-validation results of Pre-RoR with different depths.

4.3. Comparisons with state-of-the-art results of age estimation on Adience

We use 4c2f-CNN, VGG-16, Pre-ResNets and our optimal Pre-RoR-58+SD architectures with the two mechanisms to estimate age. Table 3 compares the state-of-the-art methods for age group classification on Adience dataset. We find that the accuracy increases with higher network complexity, and two mechanisms will further improve each architecture by 1.3%∼1.7%, which demonstrates the versatility of two mechanisms in different models. Particularly, our Pre-ROR-58+SD with two mechanisms obtains a single-model accuracy of 64.17±3.81% on Adience, which is the now state-of-the-art performance to our best knowledge.

The accuracy of Pre-ROR-58+SD with two mechanisms is better than 64.0±4.2% of DEX which pre-trained on ImageNet [27] and IMDB-WIKI (523,051 face images) [1]. Although DEX can achieve competitive results, it needs very large dataset IMDB-WIKI for pre-training. Our method can learn age and gender representation from scratch without the IMDB-WIKI and achieve the best performance. Our VGG-16 with two mechanisms also outperforms DEX (also based on VGG-16) which only pre-trained on ImageNet but without IMDB-WIKI. These results demonstrate that our method can improve the optimization ability of networks and alleviate over-fitting on Adience dataset. We have reason to argue that better performance can be achieved by pre-training on extra datasets.

Method	Exact Acc(%)	1-off(%)
SVM-dropout [12]	45.1±2.6	79.5±1.4
4c2f-CNN [20]	50.7±5.1	84.7±2.2
R-SAAFc2 [21]	53.5	87.9
DEX w/o IMDB-WIKI pretrain [1]	55.6±6.1	89.7±1.8
DEX w/ IMDB-WIKI pretrain [1]	64.0±4.2	96.60±0.90
4c2f-CNN	52.62±4.37	88.61±2.27
4c2f-CNN with two mechanisms	53.96±3.80	90.04±1.54
VGG-16	54.64±4.76	89.93±1.87
VGG-16 with two mechanisms	56.11±5.05	90.66±2.14
Pre-ResNets-34	60.15±3.99	90.90±1.67
Pre-ResNets-34 with two mechanisms	61.89±4.16	93.50±1.33
Pre-RoR-58+SD	62.50±4.33	93.63±1.90
Pre-RoR-58+SD with two mechanisms	64.17±3.81	95.77±1.24

Table 3. The age cross-validation results by different methods.

5. CONCLUSION

This paper proposes a Residual networks of Residual networks architecture (RoR) for high-resolution facial images age classification in the wild. Two modest mechanisms, pre-training by gender and training with weighted loss layer, are used to improve the performance of age estimation. By Pre-RoR with two mechanisms, we obtain new state-of-the-art performance on Adience dataset for age group classification in the wild.

6. REFERENCES

[1] Rasmus Rothe, Radu Timofte, and Luc Van Gool, “Deep expectation of real and apparent age from a single image without facial landmarks,” *International Journal of Computer Vision*, 2016.

[2] Asuman Gunay and Vasif V. Nabiyev, “Automatic age classification with lbp,” *International Symposium on Computer and Information Sciences*, pp. 1–4, 2008.

[3] Shuicheng Yan, Ming Liu, and Thomas S. Huang, “Extracting age information from local spatially flexible

patches," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 737–740, 2008.

[4] Guodong Guo, Guowang Mu, Yun Fu, and Thomas S. Huang, "Human age estimation using bio-inspired features," *CVPR*, pp. 112–119, 2009.

[5] Yun Fu and Thomas S. Huang, "Human age estimation with regression on discriminative aging manifold," *IEEE Transactions on Multimedia*, vol. 10, no. 4, pp. 578–584, 2008.

[6] Guodong Guo, Yun Fu, Charles R. Dyer, and Thomas S. Huang, "Image-based human age estimation by manifold learning and locally adjusted robust regression," *IEEE Transactions on Image Processing*, vol. 17, no. 7, pp. 1178–1188, 2008.

[7] Guodong Guo and Guowang Mu, "Joint estimation of age, gender and ethnicity: Cca vs. pls," *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, pp. 1–6, 2013.

[8] Andreas Lanitis, Draganova Chrisina, and Christodoulou Chris, "Comparing different classifiers for automatic age estimation," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 34, no. 1, pp. 621–628, 2004.

[9] Rasmus Rothe, Radu Timofte, and Luc Van Gool, "Morph: A longitudinal image database of normal adult age-progression," *International Conference on Automatic Face and Gesture Recognition*, pp. 341–345, 2006.

[10] Neeraj Kumar, Alexander C. Berg, Peter N. Belhumeur, and Shree K. Nayar, "Attribute and simile classifiers for face verification," *ICCV*, pp. 365–372, 2009.

[11] Andrew C. Gallagher and Tsuhan Chen, "Understanding images of groups of people," *CVPR*, pp. 256–263, 2009.

[12] Eran Eidinger, Roee Enbar, and Tal Hassner, "Age and gender estimation of unfiltered faces," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 12, pp. 2170–2179, 2014.

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, "Imagenet classification with deep convolutional neural networks," *NIPS*, pp. 1097–1105, 2012.

[14] Will Y. Zou, Xiaoyu Wang, Miao Sun, and Yuanqing Lin, "Generic object detection with dense neural patterns and regionlets," *BMVC*, 2014.

[15] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[16] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, "Going deeper with convolutions," *CoRR*, vol. abs/1409.4842, 2015.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015.

[18] Dong Yi, Zhen Lei, and Stan Z. Li, "Age estimation by multi-scale convolutional network," *ACCV*, pp. 144–158, 2014.

[19] Xiaolong Wang, Rui Guo, and Chandra Kambhamettu, "Deeply-learned feature for age estimation," *IEEE Winter Conference on Applications of Computer Vision*, pp. 534–541, 2015.

[20] Gil Levi and Tal Hassner, "Age and gender classification using convolutional neural networks," *CVPR Workshop*, pp. 34–42, 2015.

[21] Le Hou, Dimitris Samaras, Tahsin M. Kurc, Yi Gao, and Joel H. Saltz, "Neural networks with smooth adaptive activation functions for regression," *CoRR*, vol. abs/1608.06557, 2016.

[22] Ke Zhang, Miao Sun, Tony X. Han, Xingfang Yuan, Liru Guo, and Tao Liu, "Residual networks of residual networks: Multilevel residual networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. accepted, 2017.

[23] Young Ho Kwon and Niels da Vitoria Lobo, "Age classification from facial images," *CVPR*, pp. 762–767, 1994.

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Identity mappings in deep residual networks," *CoRR*, vol. abs/1603.05027, 2016.

[25] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q. Weinberger, "Deep networks with stochastic depth," *CoRR*, vol. abs/1603.09382, 2016.

[26] S. Gross and M. Wilber, "Training and investigating residual nets," *[Online]. Available:http://torch.ch/blog/2016/02/04/resnets.html*, 2016.

[27] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei Fei, "Imagenet: A large-scale hierarchical image database," *CVPR*, pp. 248–255, 2009.