Age Estimation System Using Deep Residual Network Classification Method

Arna Fariza
Informatics and Computer Eng. Dept.
Politeknik Elektronika Negeri Surabaya
Surabaya, Indonesia
arna@pens.ac.id

Mu'arifin
Informatics and Computer Eng. Dept.
Politeknik Elektronika Negeri Surabaya
Surabaya, Indonesia
muarifin@pens.ac.id

Agus Zainal Arifin
Informatics Engineering Dept.
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
agusza@cs.its.ac.id

***Abstract*— The human face has biometric properties that are important for providing age information because of the aging process of the face. Automatic Age estimation is a difficult problem because the relationship between facial images and age is not very linear. Deep residual network (Resnet) is a neural network convolution architecture that was easier to optimize and can gain accuracy results from a considerably increasing depth. In this paper, we propose a new approach age estimation on convolution neural network (CNN) using the deep residual network (Resnet) model. Through the literature, Resnet achieves superior results when compared with other state-of-the-art image classifications. We compare a new generation of deep residual network called ResNeXt with Resnet and a basic linier regression model architecture.We user UTKFace dataset to evaluate the performance of residual network for age estimation of the range 1-100 years old. The result shows that the ResNeXt-50 (32×4d) architecture achieves a better age estimation results than Resnet-50 and linier regression.**

Keywords—age estimation, human face, deep residual network, ResNeXt.

I. INTRODUCTION

Age estimation plays a significant role for individual identification in law enforcement, security control, and human social interaction. The human face has biometric properties that are important for providing age information because of the aging process of the face. Automatic age estimation becomes a challenging problem in computer vision because the aging process on the face is not only determined by intrinsic factors, e.g. genetic factors, but also by extrinsic factors, e.g. lifestyle, expression, and environment [1]. Therefore, age estimation of human faces is a difficult task because of the highly non-linear relationship between facial images and chronological age [2]. Thus, mapping pixel images to its corresponding age requires robust and accurate functions.

The most widely used method for estimating the age of facial images consists of two steps. They are local feature extraction and regression (or classification). Local feature extraction is used to get strong representations for irrelevant factors, such as identity, gender, ethnicity, poses, illuminations, expressions, and so on. Classification or regression models are used to study the closeness of features to chronological age. The K-Nearest Neighbors [3], Multilayer Perceptron [4], and Support Vector Machines (SVM) [5] classification models are most commonly used, and the regression methods that are often used to include quadratic regression [6], Support Vector Regression (SVR) [5] and multi-instance regression [7].

Deep learning techniques such as Convolutional Neural Networks (CNN) have been applied to estimation of human age to study the aging features straight from large-scale face data [8].This system leads to a fully end-to-end system that can estimate the age of pixel images directly. Experimental results indicate that the aging patterns studied in causing significant performance improvements in the benchmark data set [9].

Deep Residual Learning was introduced by He et al. [10] is a neural network convolution architecture that was easier to optimize and can gain accuracy results from a considerably increasing depth. It redefines layers as learning residual functions explicitly by referring to the input layer, instead of studying functions that are not referenced. Residual network (Resnet) with deeper depths than Visual Geometric Group (VGG) nets [11] has lower complexity and produces lower errors on the Imagenet data set. The ResNeXt architecture is an extension of the deep residual network which substitutes the standard residual block with one that leverages a "split-transform-merge" strategy [12]. It shows that even under the restricted condition of maintaining complexity, increasing cardinality can increase Imagenet classification accuracy. Moreover, increased cardinality on ResNeXt is more effective than becoming deeper or wider when capacity increases.

In this paper, we propose a new approach age estimation of face images on CNN using the deep residual network model. Through the literature, Resnet achieves superior results when compared with other state-of-the-art image classifications. We compare a new generation of deep residual network called ResNeXt with Resnet and a basic linier regression model architecture. We use UTKFace data set to evaluate the performance of residual network for age estimation of the range 1-100 years old.

II. RELATED WORK

A. Age Estimation

The previous work of age estimation was automatically carried out by Geng et al.[13] who proposed the AGES (AGing PattErn Subspace) method. These method models the pattern of aging as a sequence of faces of certain individuals in time, by creating a representative subspace. Estimates are made by projecting facial images onto aging boundaries with the best reconstruction. A person's facial features that are almost identical in several age ranges are a challenge. Because the images available to certain people are usually very limited, many researchers focus on developing an impersonal approach. For example, Local Binary Patterns [14] and learning ordinal discriminatory features [15] for classifying face images. Learning algorithms, for example, Principal Component Analysis (PCA), have been used to achieve better age estimation performance [16].

Recently, the use of the CNN method for age estimation has been widely adopted because of its superior performance compared to the statistical regression and classification

methods. Yi et al. [8] introduced a multi-task learning method with shallow CNN. Wang et al. [9] CNN trained more deeply to extract features from various layers, and these features were then integrated by PCA. Rothe et al. adopted a very deep VGG-16 architecture [17] for estimated age.

B. Convolution Neural Network

Computer vision technique introduces Convolutional Neural Network (CNN) as a development of Neural Networks. CNN achieves great successfully end-to-end classification for various tasks related to images such as image classification [18], objects for detection / localization [19], face recognition [8] [9] and image segmentation [20].

CNN with deep architecture can recognize objects in large-scale drawing data sets. Some components are used to build a more effective CNN model: the activation unit called rectified linear unit (ReLU) [21] helps accelerate convergence during training and dropout regularize prevents over fitting by setting some activation units to zero in certain layers [22].

C. Deep Residual Network

Deep Residual Network [10] is feasibly the most innovative work in computer vision in recent years. ResNet allows to train up to hundreds or even thousands of layers and still achieve interesting performance. Deep networks can cause vanishing gradient problems because gradients are propagated back to the previous layer, repetitive multiplication can create the gradient very small. As a result, when the network gets deeper, its performance becomes saturated or even declines rapidly.

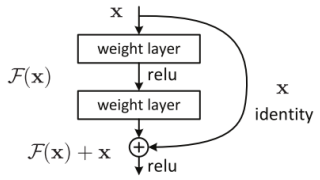

Fig. 1. A residual block [10].

The ResNet architecture introduces skip connection that allows input x to skip one or more layers without passing through some layer weights as shown in Fig. 1. Desired base mapping $H(x)$ which fit the stacked nonlinear layers mapping $F(x)$where $F(x)=H(x)-x$, will be rearranged to $F(x)+x$. When mapping x identity is optimal, this makes it easier to push the residue to zero instead of fulfilling the x identity using a stacked nonlinear layer's mapping. "Skip / shortcut connection" causes the formulation $F(x)+x$ to enter the feed forward neural network. Therefore, the residual block allows the gradient to flow through the shortcut connection to the previous layer unimpeded. This residual block explicitly causes lower training errors because it is easier to map stacked nonlinear layers mapping $F(x)$ than desired base mapping $H(x)$.

Xie et al. [12] proposed a variant of ResNet that is code named ResNeXt. It secured second place in ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2016, a competition of evaluating algorithms to achieve higher accuracy on the given data set for image classification and object recognition. It has a homogeneous and multi-branch architecture with a few hyper-parameters settings called cardinality - the number of independent paths, to provide a new way of adjusting the capacity of the model. The building block of ResNeXt can be seen in Fig.2. Similar with Resnet, it architectures to follow the split-transform-merge paradigm, unless the output of different paths is combined by adding them together. Experiments show that accuracy can be obtained more expeditiously by increasing cardinality than by becoming deeper or wider. The detailed architecture of ResNeXt-50 with cardinality C = 32 and width bottleneck d = 4 can be seen in Fig.3 while the ResNet-50 architecture which is a special case of ResNeXt-50 with C = 1, d = 64.

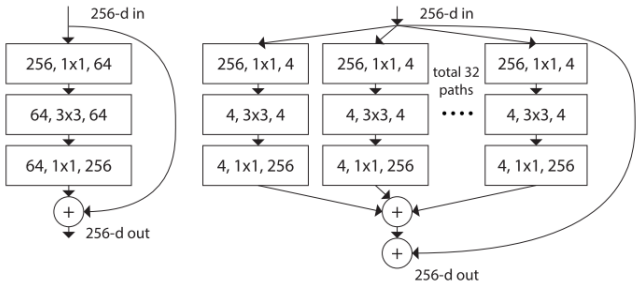

Fig. 2. (a) A building block of Resnet [10]; (b) A building block of ResNeXt with cardinality = 32 [12].

stage	output	ResNet-50	**ResNeXt-50 (32×4d)**
conv1	112×112	7×7, 64, stride 2	7×7, 64, stride 2
conv2	56×56	3×3 max pool, stride 2	3×3 max pool, stride 2
		[1×1, 64; 3×3, 64; 1×1, 256] ×3	[1×1, 128; 3×3, 128, C=32; 1×1, 256] ×3
conv3	28×28	[1×1, 128; 3×3, 128; 1×1, 512] ×4	[1×1, 256; 3×3, 256, C=32; 1×1, 512] ×4
conv4	14×14	[1×1, 256; 3×3, 256; 1×1, 1024] ×6	[1×1, 512; 3×3, 512, C=32; 1×1, 1024] ×6
conv5	7×7	[1×1, 512; 3×3, 512; 1×1, 2048] ×3	[1×1, 1024; 3×3, 1024, C=32; 1×1, 2048] ×3
	1×1	global average pool 1000-d fc, softmax	global average pool 1000-d fc, softmax
# params.		**25.5**×10^6	**25.0**×10^6
FLOPs		**4.1**×10^9	**4.2**×10^9

Fig. 3. Comparison of detailed architecture of ResNet-50 and ResNeXt-50 with 32×4d template. Inside the bracketsare show the shape of a residual block, and outside the bracket show the number of grouped convolution with 32 groups [12].

III. Residual Network Classification for Age Estimation Methodology

This section explains the methodology of classification using residual network CNN for estimate age according to `

A. Age Estimation Dataset

In this paper, we use UTKFace data set to analyze the performance of residual network for age estimation. UTKFace large-scale data collection provides more than 20,000 face images with annotations of age, gender, and ethnicity from around the world with ages ranging from 0 to 116 years. The Image sets consist of various poses, facial expressions, lighting, occlusion, resolution, use of accessories, etc. The UTKFace data are available in copy version with 224 × 224 pixels weight and height. We select 23.663 data and divide into 90% training data and 10% test data. Some examples of

face images in the UTKFace data set are shown in Fig. 4. This data set is very challenging and can serve as an excellent benchmark for evaluating the performance of the age estimation algorithm due to unequal age distribution, different image brightness and image positions. The age distribution of the training data set can be seen in Fig. 5, while the composition of training data and testing in the age range can be seen in Table I.

Fig. 4. Some examples of UTKFace image with different image brightness and image positions as a challenge of age estimation evaluation.

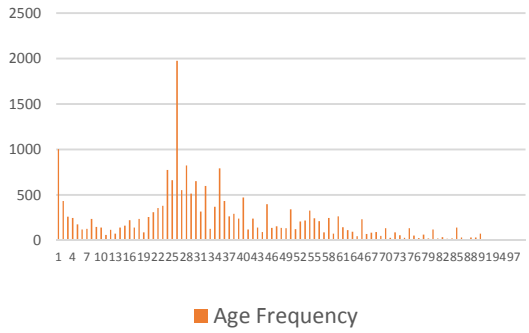

Fig. 5. Distribution frequency of training data.

B. System Design

The design of age estimation system based on Resnet architecture can be seen in Fig 2.The first step in the age estimation process is collecting data and data validation. The UTKFace data set has labeled each image, but has not been divided into a collection of training data, validation and testing. Therefore, we select data and organize one folder for each age. In this study, the training data collection contained 80% of the total data labeled.This data will be used to train machines of various types of images. The validation data set will contain 10% of the total labeled data, which tests how well our machine's performance against known labeled data.The test data set will contain 10% of the total data in a format that is not labeled. This test data is used to test how well our machine can classify data that it has never seen. Each image is normalized to obtain sharper images.

TABLE I. COMPOSITION OF TRAINING AND TESTING UTKFACE DATASET

Age	Total Data	Training Data	Test Data
1-10	3.206	2.885	321
11-20	1.658	1.493	165
21-30	7.781	6.998	783
31-40	4.336	3.901	435
41-50	2.100	1.890	210
51-60	2.209	1.998	211
61-70	1.172	1.054	118
71-80	684	615	69
81-90	451	407	44
91-100	66	60	6

In the process of deep neural networks, we determine the model used is the deep residual network (Resnet) consist of Resnet-50 and ResNeXt-50 (32×4d). Before the training, process defined the criteria for loss function, optimization and learning level. Cross-entropy loss is commonly used as a loss function for classification problems which in practice works relatively well for training machine learning models. If the model predicts the probability distribution p ($y = i$) for each class i = 1,2, ..., C, if the correct class is c, then cross-entropy loss is

$$\mathcal{L} = -\sum_{i=1}^{C} 1[i = c] \log p(y = i) = -\log p(y = c) \quad (1)$$

If there is an optimal distribution of the correct class, then a higher probability for the class will be set because it is more similar. Adam [23] is an adaptive learning rate optimization algorithm that can used instead of the classical stochastic gradient descent procedure to update network weights iterative based in training data. Adam showed outstanding performance gains in terms of training speed. The algorithm utilizes the power of the adaptive learning level method to find individual learning levels for each parameter.

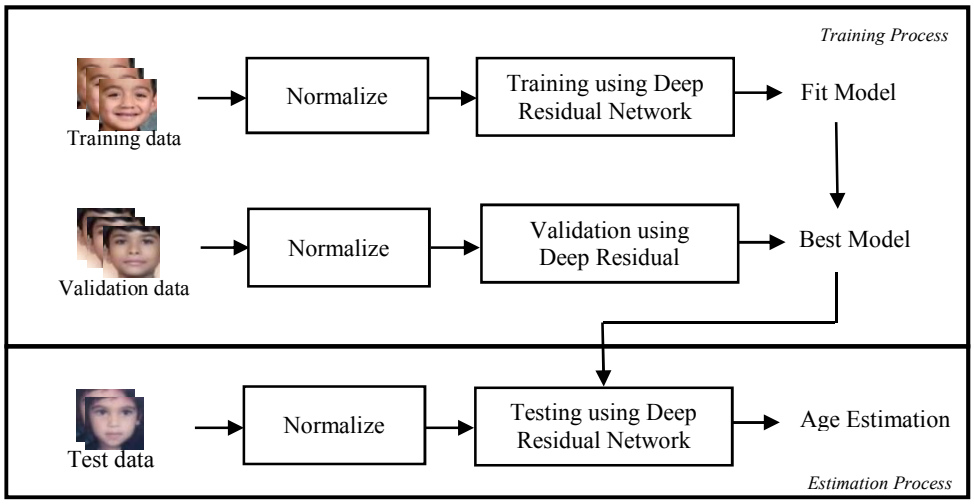

Fig. 6. System design of age estimation using deep residual networ

In the training process, a batch of images is loaded and a feed forward loop is performed. Then calculate the loss function, and use the optimizer to apply gradient descent in back-propagation. In the validation process, set the model and return to the training process after completion. After a number of epochs, the model is stored for the estimation process.

C. Evaluation Criteria

Age estimation algorithm is evaluated from the closeness between the estimated actual value. The standard metrics for testing age estimation algorithms performance is calculated using Mean Absolute Error (MAE). Smaller MAE value obtains good performance of the age estimation algorithm. MAE calculates the absolute error between estimated age and actual age as defined in equation (2).

$$MAE = \frac{\sum_{i=1}^{N} |y_i - \bar{y}_i|}{N} \quad (2)$$

where N is the number of testing samples, y_i is the ground truth age and $\bar{y}_i$ is the predicted age of the i-th sample.

IV. Experiment Result

In this section, we analyze the performance of classification method using residual network Resnet-50 and ResNeXt-50 (32×4d) compared with regression method using linier regression in aspect training and estimation accuracy. The data length of the training and testing data set is 21,301 and 2,362 respectively. The learning rate are 0.001 and the training epochs are 200.

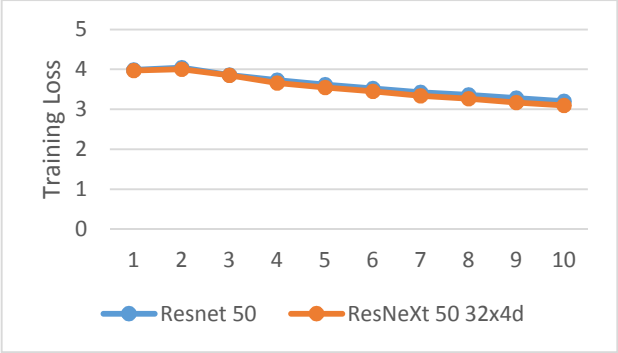

Fig. 7. Training loss comparison between Resnet-50 and ResNeXt-50 (32×4d)

Fig. 7 and Fig. 8 present the training loss and training accuracy comparison of the Resnet-50 and ResNeXt-50 (32×4d) in the 2 axis within 10 epochs. The loss value in the training and validation sets is calculated based on the residual network architecture to get the smallest loss value interpreting the best model. The loss value is the sum of the errors made by the training and validation set. While the accuracy value represents the percentage of the training set and validation that matches the label. Accuracy values close to 100 percent indicates excellent training and validation results. An accurate model can learn and improve parameters during the training process. The training set and validation are entered into the model, and the number of errors is recorded by comparing the actual labels on each epoch. The blue line represents the Resnet-50 and the red line represents the ResNeXt-50 (32×4d) of the training loss and accuracy. In these figures, ResNeXt-50 (32×4d) deliver a faster convergence speed with smaller training loss and bigger training accuracy then Resnet-50.

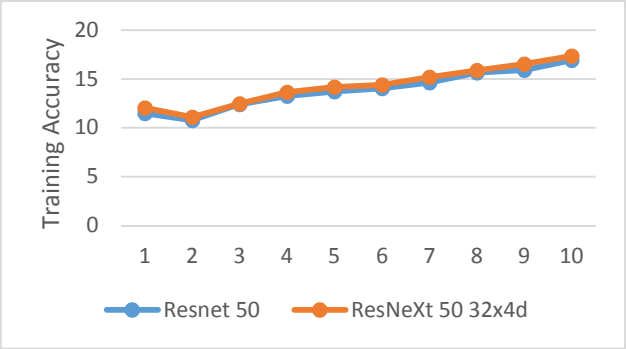

Fig. 8. Training accuracy comparison between Resnet-50 and ResNeXt-50 (32×4d)

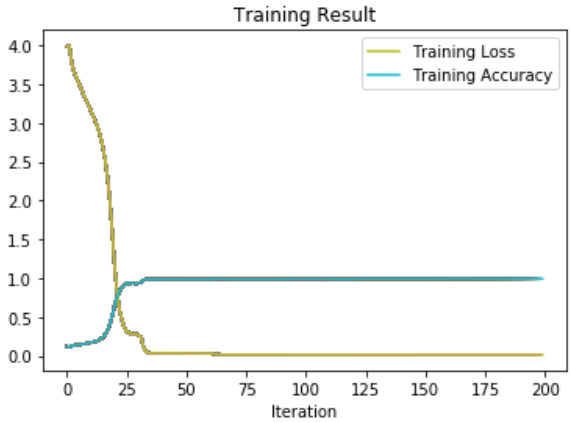

Fig. 9. Training result of Resnet 50

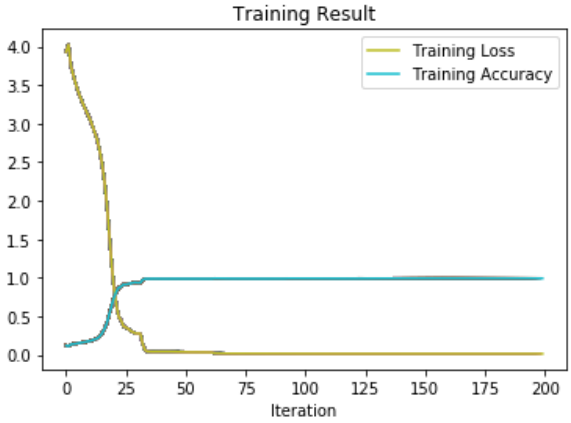

Fig. 10. Training result of ResNeXt-50 (32×4d)

The training graph of Resnet-50 run in 200 epochs gets the training loss minimal 0.012, the training accuracy 99.15%, while ResNeXt-50 (32×4d) gets the training loss minimal 0.012, the accuracy 99.15% as shown in Fig. 9 and Fig.10. We also compare the training results with linier regression model, but it converges slower and cannot reach better accuracy. The training graph shows that in 1000 epoch, it gets the training loss minimal 1.2807 and the training accuracy 64.37% as shown in Fig 11. Overall it shows that

Resnet-50 and ResNeXt-50 (32×4d) are feasible for this age estimation application, compared with linier regression.

Furthermore, we evaluate the MAE of a testing result. We use the model from the training results to test the unlabeled testing dataset. Table II shows ResNeXt-50 (32×4d) obtain the better MAE than Resnet-50, even though the training process reaches the same training loss and training accuracy. ResNeXt-50 (32×4d) reduces MAE by 1.53 years from Resnet-50 significantly. Likewise, Resnet-50 and ResNeXt-50 (32×4d) show better estimation results than linear regression. Since the UTKFace data set is built from faces in the wild, those methods organized experiments on this challenging data set. Overall, it shows that the deep residual network classification is appropriate to solve the problem of age estimation.

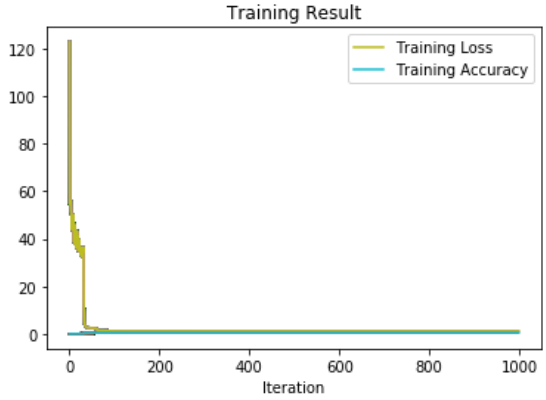

Fig. 11. Training result of regression logistic

TABLE II. COMPARISON OF ESTIMATION RESULT

Method	Training Loss	Training Accuracy	MAE of Estimation
Linier Regression	1.2807	64.37%	11.73
Resnet-50	0.012	99.15%	8.74
ResNeXt-50 (32×4d)	0.012	99.15%	7.21

V. CONCLUSION

The results of age estimation experiments with deep residual network ResNeXt-50 (32×4d) show superior results than Resnet-50 and linear regression. In general, we obtain that the deep residual network classification is appropriate to solve the regression problem of age estimation system based on face recognition.

In our future work, we add classification model to improve the residual network architecture. We also add the scheduled learning rate to increase the training performance. With these strategies, the estimating error can reduce significantly.

ACKNOWLEDGMENT

We would like to thank to Politeknik Elektronika Negeri Surabaya who has supported this research.

REFERENCES

[1] H. Han, C. Otto, and A.K. Jain, "Age estimation from face images: Human vs. machine performance", In 2013 International Conference on Biometrics (ICB), 2013, pp. 1-8.

[2] D. Yi, Z. Lei, and S.Z. Li, (2014, November). "Age estimation by multi-scale convolutional network", In Asian conference on computer vision, 2014, pp. 144-158.

[3] A. Gunay and V.V. Nabiyev, "Automatic age classification with LBP", In 2008 23rd International Symposium on Computer and Information Sciences, 2008, pp. 1-4

[4] A. Lanitis, C. Draganova, and C. Christodoulou, "Comparing different classifiers for automatic age estimation," IEEE Trans. Syst., Man, Cybern. B, Cybern., vol. 34, no. 1, pp. 621–628, Feb. 2004.

[5] G. Guo, G. Mu, Y. Fu, and T. S. Huang, "Human age estimation using bioinspired features," in Proc. IEEE Conf. Comput. Vision Pattern Recognit., 2009, pp. 112–119.

[6] G. Guo, Y. Fu, C. R. Dyer, and T. S. Huang, "Image-based human age estimation by manifold learning and locally adjusted robust regression," IEEE Trans. Image Process., vol. 17, no. 7, pp. 1178–1188, Jul. 2008.

[7] B. Ni, Z. Song, and S. Yan, "Web image and video mining towards universal and robust age estimator," IEEE Trans. Multimedia, vol. 13, no. 6, pp. 1217–1229, Dec. 2011.

[8] D. Yi, Z. Lei, and S. Z. Li, "Age estimation by multi-scale convolutional network," in Proc. Asian Conf. Comput. Vision, 2015, pp. 144–158.

[9] X. Wang, R. Guo, and C. Kambhamettu, "Deeply-learned feature for age estimation," in Proc. IEEE Winter Conf. Appl. Comput. Vision, 2015, pp. 534–541.

[10] K. He, X. Zhang, S. Ren, and J. Sun, " Deep residual learning for image recognition", In Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770-778.

[11] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition", InICLR, 2015.

[12] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks", In Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1492-1500.

[13] X. Geng, Z.-H. Zhou, and K. Smith-Miles, "Automatic age estimation based on facial aging patterns," IEEE Trans. Pattern Anal. Mach. Intell., vol. 29, no. 12, pp. 2234–2240, Dec. 2007.

[14] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol. 28, no. 12, pp. 2037–2041, Dec. 2006.

[15] C. Li, Q. Liu, J. Liu, and H. Lu, "Learning ordinal discriminative features for age estimation," in Proc. IEEE Conf. Comput. Vision Pattern Recognit.,2012, pp. 2570–2577.

[16] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. NewYork, NY, USA: Wiley, 2012.

[17] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. 25th Int. Conf. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[19] R. Girshick, J. Donahue, T. Darrell, and J.Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in Proc. IEEE Conf. Comput. Vision Pattern Recognit., 2014, pp. 580–587.

[20] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in Proc. IEEE Conf. Comput. Vision Pattern Recognition., 2015, pp. 3431–3440.

[21] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in Proc. 27th Int.Conf.Mach. Learn., 2010, pp. 807–814.

[22] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," J. Mach. Learn. Res., vol. 15, no. 1, pp. 1929–1958, 2014.

[23] D.P. Kingma, and J. Ba, "Adam: A method for stochastic optimization", arXiv preprint arXiv:1412.6980, 2014.