# Smart Facial Age Estimation with Stacked Deep Network Fusion

Kuan-Hsien Liu[1], Pak Ki Chan[2], and Tsung-Jung Liu[2]
[1]*National Taichung University of Science and Technology, Taichung*
[2]*National Chung Hsing University, Taichung*

***Abstract*--In this paper, we proposed a stacked deep network fusion model for the apparent human facial age estimation. Three well-performed deep architectures are utilized in the first three stages and the estimation results for each architecture are fused in the last stage to boost the overall performance. In the first stage, a pre-trained deep CNN model is fine-tuned for the gender classification task. For the second stage, two gender-specific age groupers are built to classify the facial images into two non-overlapping age groups. In the third stage, ages are estimated from the three deep networks and fed to the fuser of the last stage to refine age estimation results. Experimental results demonstrate a significant performance improvement of the proposed approach over the state-of-the-art deep CNN models.**

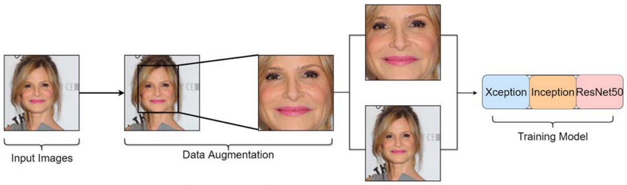

Fig. 1. Data augmentation procedure.

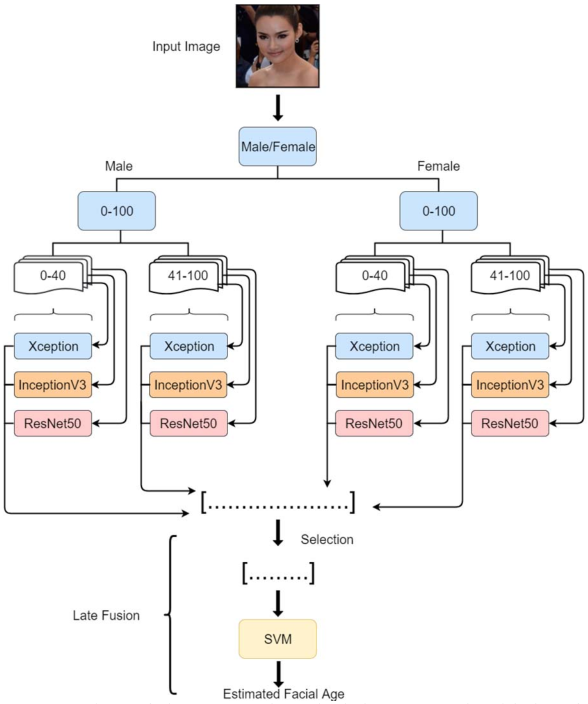

Fig. 2. Overview of the proposed stacked deep network with late fusion framework for smart facial age estimation, where blue blocks use Xception.

## I. INTRODUCTION

Over the past years, Facebook users have been uploading 350 Million new photos each day, and the upload of different images to the Internet and social network has shown an explosive increase. This rapid data growth rate has made the deep learning based approaches feasible for tackling various problems in computer vision and other fields. Among all the related problems, such as image retrieval, object recognition, multimedia quality assessment [9], human facial age estimation remains an active and challenging task.

Some recent age estimation methods [4, 7, 8] have shown good results and some [4] has been developed based on deep convolutional neural networks (CNN). These deep model based age estimation methods show good results and motivate us to design a novel age estimation framework based on the best-performed deep models.

## II. PROPOSED APPROACH

In this work, we proposed a novel approach for apparent age estimation, which includes facial data augmentation in the preprocessing step. In general, a common problem for some image datasets and video datasets would be inadequate amount of data or unbalanced data distribution that could cause the lack of data for model training. To resolve this issue, we use traditional data augmentation consisting of use of general affine transformations to the training data. For each input image, we generate a companion image that extracted the area of human face from corresponding image by face recognition library of the Adam Geitgey. Then, both original image and augmented image are fed into the proposed framework. In this way, we not only double the size of the dataset, but also focus more on age related region (i.e., face). The data augmentation procedure is illustrated in Fig. 1.

In the first stage of stacked model, an ImageNet pre-trained model, Xception [1], is fine-tuned on the facial age dataset to conduct gender grouping. The training data are divided into male and female groups, and gender-grouped training data is used for learning this Xception-based gender grouper.

For the second stage of stacked network, two Xception models pre-trained on ImageNet are fine-tuned on the male and female face images of the same dataset, respectively. The male training facial images are divided into two non-overlapping age groups, where one group consists of face images from 0 to 40 years old, and another one has face images ranging from 41 to 100 years old. The male age grouper is learned based on the above pre-defined two age groups of male training data. The female age grouper is also built in the same fashion.

In the third stage of the proposed framework, three state-of-the-art deep CNN architectures, Xception [1], Inception-V3 [2], and ResNet-50 [3], are employed to conduct age estimation for each of the two age groups. For the male training data, in the 0-40 year-old age group, three deep age estimators based on the Xception, Inception-V3, and ResNet-50 are individually trained

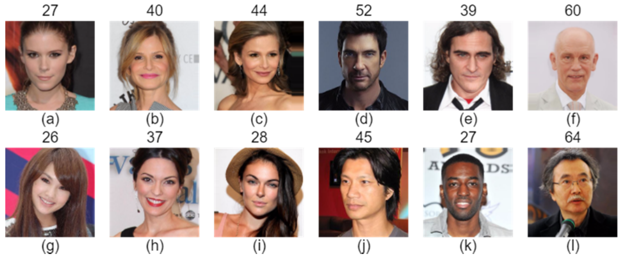

Fig. 3. Some sample images from IMDB with (gender, age): (a) F(female), 27, (b) F, 40, (c) F, 44, (d) M(male), 52, (e) M, 39, (f) M, 60. Sample images from WIKI: (g) F, 26, (h) F, 37, (i) F, 28, (j) M, 45, (k) M, 27, (l) M, 64.

to predict ages within this group. Another three deep estimators with the same structure are also individually learned to predict ages within the 41-100 year-old age group. We do the same age estimation procedure for the female training set. Each of three deep models runs 60 iterations and hence 60 results are generated for each deep model. At the output of this stage, 180 age estimation results can be obtained for each *gender-age group-deep estimator* model and will be sent to the last stage for fuser inputs.

For the fourth stage, which is also the last stage of the proposed system, we use a sequential selection algorithm to suboptimally combine some age estimation results from the 180 estimated ages to form a subset. For the consideration of simplicity and efficiency, each formed subset is trained sequentially using a support vector machine (SVM) with RBF kernel and the subset with the best mean absolute error (MAE) performance is chosen for further fusion. The proposed framework is also depicted in Fig. 2 for clear reference.

## III. EXPERIMENT

In the experiment, the benchmark IMDB-WIKI face dataset [4] is used to validate our proposed approach. So far this is the largest publicly available dataset of face images with gender and age (0-100 years old) labels for 500K+ face images. The IMDB dataset (Fig. 3) contains the most popular 100,000 actors listed on the IMDb website and all images includes date of birth, name, and gender. The WIKI dataset (Fig. 3) was constructed by crawling images from Wikipedia website with people profile and meta information (the date when the photo was taken). The IMDB-WIKI dataset contains 460,723 face images of 20,284 celebrities from IMDb and 62,328 face images from Wikipedia, and thus 523,051 face images in total.

The experimental setting is illustrated as follows. The IMDB dataset (171,852 images) is used for training and WIKI dataset (38,138 images) is for testing. We fine-tuned ImageNet pre-trained deep models, and used Adamax[5] with a learning rate of 0.0002 for 60 training epochs and a batch size of 32 images. Finally, we use SVM [6] for the late fusion.

From Table I, we can see that using only 4 age results in single-model fusion (XXX fusion, XXI fusion, and XXR fusion) can achieve the optimal performance, while 12 age results are required for three-model fusion to obtain satisfying performance. For performance comparison, from Table II we can find that fusing 3 deep models (i.e., XXX-XXI-XXR fusion) has the best MAE performance compared with single-model fusion. However, three individual single-model fusions still have better MAE performance than the three best-performed deep neural networks.

## IV. CONCLUSION

We proposed a smart age estimation system, which is based on stacked deep neural networks with late fusion. The model performs gender classification first, and then separates the face images into two disjoint age groups. Within each age group, three deep networks are trained to be able to predict the age of facial image. Lastly, the SVM is adopted to fuse a portion of the predicted results from three deep models. The experiments on the largest publicly available database show the superior performance of age estimation.

TABLE I
MAE RESULTS OF DIFFERENT MODELS FUSION

| Net(s) used for fusion | Selected subset | MAE (years) |
|---|---|---|
| XXX fusion | X(2,3,4,7) | 6.6227 |
| XXI fusion | I(2,3,4,5) | 6.7218 |
| XXR fusion | R(2,3,4,9) | 6.6680 |
| XXX-XXI-XXR fusion | X(2,3,4,7) I(2,3,4,5) R(2,3,4,9) | 5.9672 |

X: Xception, I: Inception-V3, R: ResNet-50. XXX, XXI, and XXR each have 60 results for selection. XXX-XXI-XXR has 180 results for selection.

TABLE II
MAE COMPARISONS OF DIFFERENT APPROACHES

| Methods | MAE (years) |
|---|---|
| Xception [1] | 6.9270 |
| InceptionV3 [2] | 7.0101 |
| ResNet50 [3] | 6.9843 |
| XXX fusion [ours] | 6.6227 |
| XXI fusion [ours] | 6.7218 |
| XXR fusion [ours] | 6.6680 |
| XXX-XXI-XXR fusion [ours] | 5.9672 |

## ACKNOWLEDGMENT

The work was supported in part by the Ministry of Science and Technology, R.O.C., under Grant No. MOST 106-2221-E-005-090 and MOST 106-2218-E-025-001-MY2.

## REFERENCE

[1] F. Chollet. Xception: Deep Learning with Depthwise Separable Convolutions. In CVPR 2017.
[2] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna. Rethinking the Inception Architecture for Computer Vision. In CVPR 2016.
[3] K. He, X. Zhang, S. Ren and J. Sun. Deep Residual Learning for Image Recognition. In CVPR 2016.
[4] R. Rothe, R. Timofte and L. V. Gool. DEX: Deep EXpectation of apparent age from a single image. In ICCV 2015.
[5] D. Kingma, J. Ba. Adam: A Method for Stochastic Optimization. In ICLR 2015.
[6] C.-C. Chang and C.-J. Lin, LIBSVM: a library for support vector machines. ACM TIST, 2:27:1--27:27, 2011.
[7] K.-H. Liu, S. Yan, and C.-C. J. Kuo, "Age estimation via grouping and decision fusion," IEEE TIFS, vol. 10, no. 11, pp. 2408–2423, 2015.
[8] T.-J. Liu, K.-H. Liu, H.-H. Liu, and S.-C. Pei, "Age estimation via fusion of multiple binary age grouping systems," In IEEE ICIP, 2016.
[9] T.-J. Liu, K.-H. Liu, J. Y. Lin, W. Lin, and C.-C. J. Kuo, "A paraboost method to image quality assessment," IEEE TNNLS, vol. 28, no. 1, pp. 107–121, 2017.