FUSION NETWORK FOR FACE-BASED AGE ESTIMATION

Haoyi Wang[1] *Xingjie Wei*[2] *Victor Sanchez*[1] *Chang-Tsun Li*[1,3]

[1]Department of Computer Science, The University of Warwick, Coventry, UK
[2]School of Management, University of Bath, Bath, UK
[3]School of Computing & Mathematics, Charles Sturt University, Wagga Wagga, Australia
{h.wang.16, vsanchez, C-T.Li}@warwick.ac.uk, x.wei@bath.ac.uk

ABSTRACT

Convolutional Neural Networks (CNN) have been applied to age-related research as the core framework. Although faces are composed of numerous facial attributes, most works with CNNs still consider a face as a typical object and do not pay enough attention to facial regions that carry age-specific feature for this particular task. In this paper, we propose a novel CNN architecture called Fusion Network (FusionNet) to tackle the age estimation problem. Apart from the whole face image, the FusionNet successively takes several age-specific facial patches as part of the input to emphasize the age-specific features. Through experiments, we show that the FusionNet significantly outperforms other state-of-the-art models on the MORPH II benchmark.

Index Terms— Age Estimation, Soft Biometrics, Feature Extraction, Convolutional Neural Network

1. INTRODUCTION

Face-based age estimation is an active research topic, which is intended to predict the age of a subject based on the appearance of his or her face. Recently, Convolutional Neural Networks (CNN) have been proved to be capable of dramatically boosting the performance of many mainstream computer vision problems [9, 10, 17].

Neuroscience shows that when the primate brain is processing the facial information, different neurons respond to different facial features [1]. Inspired by this fact, we intuitively assume that the accuracy of age estimation may be largely improved if the CNN could learn from age-specific patches. Consequently, in this paper, we propose the Fusion Network (FusionNet), a novel CNN architecture for face-based age estimation. Specifically, FusionNets take the face and several age-specific facial patches as successive inputs. This data feeding sequence is shown in Fig. 1. As illustrated in the figure, there are a total of $n+1$ inputs (one face and n facial patches) being fed into the network. The aligned face, which provides most of the information, is the primary input that is fed to the lowest layer to have the longest learning path. After all the inputs are fed into the network, the final prediction is calculated based on this fused information that is learned through the convolutional layers. We show later that the input at the middle-level layers can be viewed as shortcut connections that boost the flow of the age-specific features.

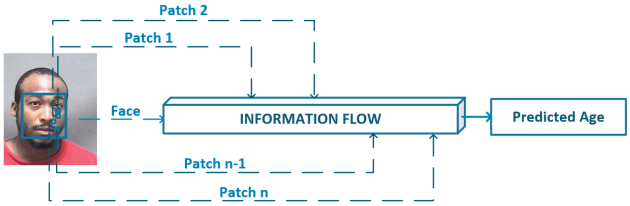

Fig. 1. Data feeding sequence in the FusionNet. The model takes the original face and a total of n facial patches as inputs.

The main contribution of this work is that we propose the FusionNet to solve the face-based age estimation problem. To the best of our knowledge, our network is the first CNN-based model in which the learning of age-specific features is enhanced by using selected input patches. Moreover, those input patches form short-cut connections that complement the learning process, which is useful to boost the performance. Experiments prove that the FusionNet significantly outperforms other state-of-the-art methods on the MORPH II benchmark [18].

The rest of this paper is organized as follows. In Section 2, we review the related works on age estimation and CNNs. In Section 3, we present the proposed method in detail. In Section 4, we show the performance of the proposed network and compare it with the results from other papers. Finally, we conclude our work in Section 5.

2. RELATED WORK

In the past few decades, many works have been conducted on face-based age estimation. One of the earliest can be traced back to [14], in which the researchers classify faces into three age groups based on the cranio-facial development theory and wrinkle analysis. Later, several popular age estimation methods are proposed, like [15, 5, 6].

Ever since Krizhevsky *et al.* [13] adopted a CNN for massive-scale image classification applications, CNNs have been applied to various computer vision problems with superior performance. With the growing size of age-related datasets, researchers have begun to use CNNs as the feature extractor. Yi *et al.* [21] propose a multi-column CNN, which is one of the earliest works that apply CNNs to age estimation. Moreover, Han *et al.* [7] have used a modified AlexNet [13] to construct a multi-task learning model for age estimation.

Most recently, Niu *et al.* [16] treat the age estimation as an ordinal regression problem. In their work, a CNN with multiple binary outputs is constructed, and each binary output solves a binary classification sub-problem with respect to the corresponding output label. Most recently, Chen *et al.* [3] also consider the ordinal relationship between different ages and proposed the Ranking-CNN for facial age estimation.

In this work, we tackle the age estimation problem from a different point of view by focusing on the representation learning. In other words, we modify the network structure to extract more representative feature by paying attention to information-rich regions.

3. FUSION NETWORK

The proposed method consists of three components, the facial patch selection, the convolutional network and the age regression. The facial patch selector is based on the Bio-inspired Features (BIF) [6] and the AdaBoost algorithm. Selected patches are subsequently fed into the convolutional network, in a sequential manner, together with the face. The final prediction is calculated based on the output of the network by using a regression method.

3.1. Facial Patch Selection

We use the BIF [6] to extract age-specific feature from aligned faces. Faces are convolved with a bank of Gabor filters [4], which can be formulated as:

$$G(x,y) = \exp(-\frac{(x'^2+\gamma^2 y'^2)}{2\sigma^2}) \times \cos(2\pi\frac{x'}{\lambda}) \qquad (1)$$

where (x, y) are the spatial coordinates, and $x' = x\cos\theta + y\sin\theta$ and $y' = -x\sin\theta + y\cos\theta$ denote the orientation of the filters with the angle $\theta \in [0, \pi]$. γ, σ, and λ are the parameters of the filters. We convolve each face with a total of 8 bands and 8 orientations of Gabor filters to generate a k-dimensional feature vector. In our experiments, k is greater than 10,000 with each element encoding one potential input for the subsequent CNN. Since we cannot use this high-dimensional feature vector in the feeding sequence directly, we need to select k' features from the BIF feature vector to form a subset where $k' << k$. We experimentally set k' to 1000 and use the top 5 most informative features as the input to the subsequent network to keep a balance between the training time and the performance. The top 5 selected features are represented as the 5 patches marked in the face in Fig. 2.

The multi-class AdaBoost is used to select the subset k' from the high-dimensional feature vector. A Decision Tree is built as the weak classifier in AdaBoost, which is similar to the implementation in [8]. Briefly, for a dataset with m samples, we pick the k' most informative features from a k-dimensional vector by using the weak classifier h,

$$\mathcal{F}_j = argmin_k(\sum_{i=1}^{m} w_i^{k'} e(h_k(x_i), y_i)) \qquad (2)$$

where $\mathcal{F}_j$ is the j-th selected feature and $j \in [1, k']$. x_i is the high dimensional feature vector after the i-th sample is filtered by Gabor filters and y_i is the associated age label. In addition, $w_i^{k'}$ is the weight in AdaBoost, which is updated and normalized after each $\mathcal{F}_j$ is found. The error function $e(h_k(x_i), y_i)$ in Eq. (2) is defined as follows:

$$e(h_k(x_i), y_i) = \begin{cases} 0 & h_k(x_i) = y_i \\ 1 & otherwise \end{cases} \qquad (3)$$

Through extensive experiments, we conclude that a 28-level Decision Tree should be implemented as the weak classifier in our case to strike a good balance between the training time and the ability to capture information.

3.2. Network Architecture

The architecture of the FusionNet is illustrated in Fig. 2. In the figure, the block arrows indicate the feature extraction process and the dashed lines between blocks denote copying. All of the blocks shown in Fig. 2 are residual blocks [9], and each block after concatenation (B1 to B5) contains bottleneck layers. Note that we do not apply feature reduction to B5 in Fig. 2, since we have found that lowering the number of feature maps right before the global pooling largely reduces the performance. Moreover, we apply a batch normalization layer [11] before each convolutional layer to improve the training speed and overall accuracy. After the convolutional stage, a global average pooling layer and a fully-connected (FC) layer are attached to generate the final output of the network.

Instead of training separate shallow CNNs for each input and concatenating the information before the final fully-connected layer, we merge the features in the convolution stage. In the FusionNet, all the features from different inputs have a longer and more efficient learning path compared to the multi-path CNN in [21]. Moreover, the common age-specific features among the inputs can be extracted and emphasized. For example, the skin feature, which has ordinal relationship to the age, can be enhanced since all the simultaneous inputs share almost the same skin texture.

The use of concatenation is inspired by the DenseNet [10]. In a DenseNet, the network is divided into several dense

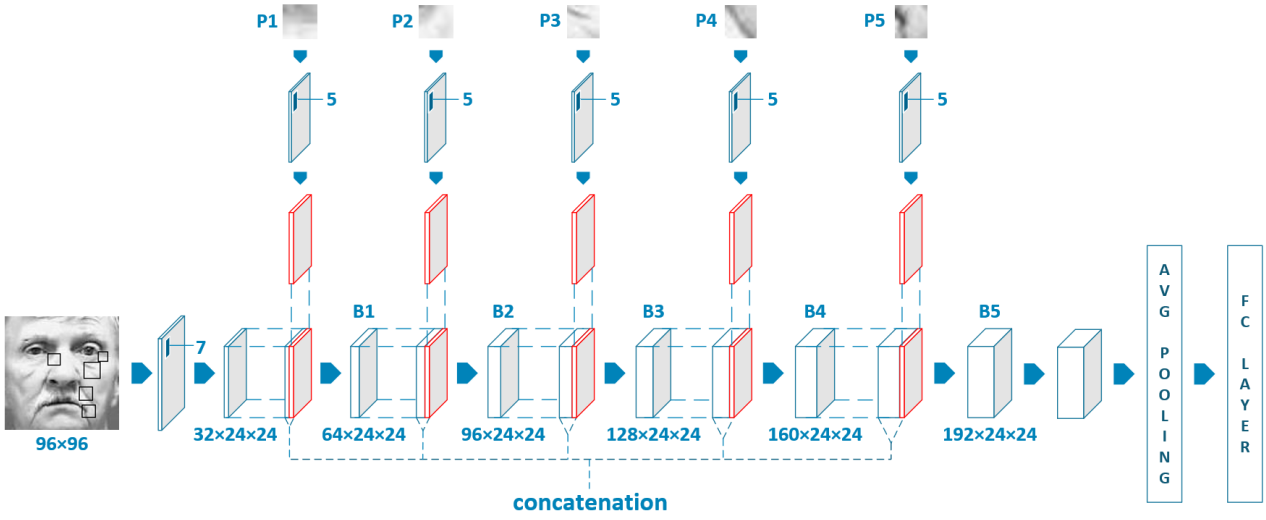

Fig. 2. The architecture of the Fusion Network for face-based age estimation. The selected patches are fed to the network sequentially as the secondary learning source. The input of patches can be viewed as shortcut connections to enhance the learning of age-specific feature. We use five patches (P1 to P5) to keep the balance between the training efficiency and the performance. The final output is produced by a single fully-connected (FC) layer.

blocks, and layers within the same block typically share the identical spatial dimension. More importantly, inside each dense block, the output of each layer flows directly into all of the subsequent layers. As a result, the l-th layer receives feature maps from all the previous layers within the same block as the input:

$$x_l = H_l([x_0, x_1, ..., x_{l-1}]) \tag{4}$$

where x represents the output of each layer and H_l denotes the learning hypothesis of the l-th layer. $[\cdot]$ is used to represent the concatenation operation.

In the FusionNet, the formulation is based on blocks, and the output of each residual block after concatenations can be represented as:

$$x_i = B_i([x_{i-1}, s_i]) \tag{5}$$

where $B_i[\cdot]$ denotes the synthesized learning function of the i-th block and $i \in [1, 5]$ since we decide to use 5 input patches in our network. Therefore, the shortcut connections in FusionNet are block-wise operations rather than layer-wise operations as in [10]. In addition, x_{i-1} is the output from the previous residual block and s_i is the feature map learned from the i-th input patch. Since the patches share common features with the original face, and based on Eq. (4) and (5), the incoming patches can be viewed as shortcut connections that refresh and amplify the flow of age-specific information.

3.3. Age Regression

Based on the fact that the discretization error becomes smaller for the regressed signal when the number of classes becomes larger [19], we calculate the final prediction through a regression approach.

After the features are processed by the fully-connected layer, we first eliminate all the negative values in the output vector and feed it to a Softmax function to form a probability distribution. Then, we normalize the distribution to make it sum up to 1.

The final prediction is the summation of products of the probabilities by the corresponding age labels.

$$\mathbf{E}(O) = \sum_{i=1}^{j} p_i y_i \tag{6}$$

where p_i denotes the normalized probability for the i-th class, y_i is the associated age label, and j is the number of classes.

4. EXPERIMENTS

4.1. Experimental Settings

We use the most frequently used MORPH II benchmark [18] for age estimation to test the performance of our network. The MORPH II dataset contains more than 55,000 facial images from about 13,000 subjects with ages ranging from 16 to 77 and an average age of 33. Following the previous works [16, 19, 2], in this work, the dataset is randomly divided into two parts, about 80% for training and the other 20% for testing. There is no overlap between the training and testing sets. To perform statistical analysis, we use 20 different partitions (with same ratio but different distribution) in the experiment and report the mean values.

Table 1. Comparison between our proposed network and a baseline model. The best result is highlighted in **bold**.

Methods	CS(n=1)	CS(n=2)	CS(n=3)	CS(n=4)	CS(n=5)	CS(n=6)	CS(n=7)	CS(n=8)
baseline	30.06%	51.07%	63.51%	74.00%	82.70%	88.07%	92.45%	95.04%
FusionNet + FAttrs + Cls	29.94%	50.51%	63.02%	73.26%	82.02%	87.50%	91.94%	94.96%
FusionNet + AdaP + Cls	**31.22%**	51.72%	67.24%	78.40%	85.26%	90.55%	93.57%	96.01%
FusionNet + AdaP + Reg	30.96%	**53.07%**	**68.35%**	**79.59%**	**86.16%**	**91.00%**	**93.97%**	**96.37%**

Table 2. MAE values of three state-of-the-art CNN-based models and our method on MORPH II dataset. The best result is highlighted in **bold**.

Methods	MAE
OR-CNN [16]	3.27
DEX [19]	3.25
Ranking-CNN [3]	2.96
baseline	3.05
FusionNet + FAttrs + Cls	3.18
FusionNet + AdaP + Cls	2.95
FusionNet + AdaP + Reg	**2.82**

We use the open-source computer vision library dlib [12] for the image preprocessing in our work. All the faces are cropped to 96×96 pixels and converted to gray-scale images since the MORPH II dataset suffers from the color cast issue. After the facial patches are selected, the cropped patches are then resized to 24×24 pixels.

The proposed network is implemented based on the open-source deep learning framework Pytorch and trained with the Stochastic Gradient Descent (SGD) algorithm with momentum. The batch size is set to 64. We train our network for 200 epochs with an initial learning rate of 0.1. The learning rate drops by a factor of 0.1 after every 50 epochs.

4.2. Results

There are two common used metrics to evaluate the performance for age estimation models, Mean Absolute Error (MAE) and Cumulative Score (CS). The MAE simply measures the average absolute difference between the predicted age and the ground truth:

$$MAE = \frac{\sum_{i=1}^{M} e_i}{M} \tag{7}$$

where e_i is the absolute error between the predicted age $\hat{l_i}$ and the input label l_i for the i-th sample. The denominator *M* is the total number of testing samples. On the other hand, the CS measures the percentage of images that are correctly classified in a certain range as:

$$CS(n) = -\frac{M_n}{M} \times 100\% \tag{8}$$

where M_n is the number of images whose predicted age $\hat{l_i}$ is in the range of $[l_i - n, l_i + n]$, and n indicates the number of years.

To demonstrate the efficiency of our proposed network, we use the CS criteria to evaluate the performance of the FusionNet compared with a baseline model, which is a plain network with all selected patches removed. In Table 1, the model in the second row represents a FusionNet taking major facial attributes like the eyes, the nose and the mouth as secondary inputs and using classification method to calculate the predicted age. The model in third row uses age-specific patches and the model in the last row uses regression to produce the final age. The reason why the second row (FusionNet + FAttrs + Cls) performs worse compared to the baseline may due to that major facial attributes carry identity-specific details rather than age-specific features, which could be treated as noise during training and degrade the performance.

We compare our approach with other recent state-of-the-art CNN-based models: DEX [19], OR-CNN [16], and Ranking-CNN [3]. To have a fair comparison, only works with the same data partition ratio are evaluated. In [19], authors use a pre-trained VGG-16 [20] as the core model and further fine-tune it on the IMDB-WIKI dataset [19]. In the comparison, we use the result without fine-tuning on the additional dataset. As shown in Table 2, the FusionNet achieves the lowest MAE of 2.82, which significantly outperforms other state-of-the-art models. This result shows that our network has a much more efficient feature extraction architecture. Moreover, the modern network design philosophy used (i.e., the residual blocks and bottleneck layers) helps to improve the performance even further.

5. CONCLUSION

In this paper, we presented the FusionNet to tackle the face-based age estimation problem. Our model takes not only the face but also other age-specific facial patches as inputs. The input facial patches can be considered as being shortcut connections in the network, which amplify the learning efficiency for age-specific features. Experiments show that our network significantly outperforms other CNN-based state-of-the-art methods on the MORPH II benchmark. In the future, we will optimize our approach by considering the ordinal and correlative relationship between ages to make more precise predictions.

6. REFERENCES

[1] L. Chang and D. Y. Tsao. The code for facial identity in the primate brain. *Cell*, 169(6):1013–1028, 2017.

[2] K. Chen, S. Gong, T. Xiang, and C. Change Loy. Cumulative attribute space for age and crowd density estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2467–2474, 2013.

[3] S. Chen, C. Zhang, M. Dong, J. Le, and M. Rao. Using ranking-cnn for age estimation. In *Computer Vision and Pattern Recognition, IEEE Conference on*, July 2017.

[4] D. Gabor. Theory of communication. part 1: The analysis of information. *Journal of the Institution of Electrical Engineers-Part III: Radio and Communication Engineering*, 93(26):429–441, 1946.

[5] X. Geng, Z.-H. Zhou, and K. Smith-Miles. Automatic age estimation based on facial aging patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(12):2234–2240, 2007.

[6] G. Guo, G. Mu, Y. Fu, and T. S. Huang. Human age estimation using bio-inspired features. In *Computer Vision and Pattern Recognition, IEEE Conference on*, pages 112–119, 2009.

[7] H. Han, A. K. Jain, S. Shan, and X. Chen. Heterogeneous face attribute estimation: A deep multi-task learning approach. *arXiv preprint arXiv:1706.00906*, 2017.

[8] H. Han, C. Otto, X. Liu, and A. K. Jain. Demographic estimation from face images: Human vs. machine performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(6):1148–1161, 2015.

[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition, IEEE Conference on*, pages 770–778, 2016.

[10] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Computer Vision and Pattern Recognition, IEEE Conference on*, July 2017.

[11] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456, 2015.

[12] D. E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[14] Y. H. Kwon and N. da Vitoria Lobo. Age classification from facial images. In *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, pages 762–767, 1994.

[15] A. Lanitis, C. J. Taylor, and T. F. Cootes. Toward automatic simulation of aging effects on face images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(4):442–455, 2002.

[16] Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua. Ordinal regression with multiple output cnn for age estimation. In *Computer Vision and Pattern Recognition, IEEE Conference on*, pages 4920–4928, 2016.

[17] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *British Machine Vision Conference*, 2015.

[18] K. Ricanek and T. Tesafaye. Morph: A longitudinal image database of normal adult age-progression. In *Automatic Face & Gesture Recognition, IEEE International Conference on*, pages 341–345, 2006.

[19] R. Rothe, R. Timofte, and L. Van Gool. Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision*, pages 1–14, 2016.

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

[21] D. Yi, Z. Lei, and S. Z. Li. Age estimation by multi-scale convolutional network. In *Asian Conference on Computer Vision*, pages 144–158. Springer, 2014.