Comparative Convolutional Neural Network for Younger Face Identification

Liangliang Wang, Deepu Rajan*

***Abstract*—We consider the problem of determining whether a pair of face images can be distinguishable in terms of age and if so, which is the younger of the two. We also determine the degree of distinguishability in which age differences are categorized into large, medium, small and tiny. We propose a comparative convolutional neural network combining two parallel deep architectures. Based on the two deep learnt face features, we introduce a comparative layer to represent their mutual relationships, followed by a concatenatation implementation. Softmax is adopted to complete the classification task. To demonstrate our approach, we construct a very large dataset consisting of over 1.7 million face image pairs with young/old labels.**

***Index Terms*—Younger face identification, comparative representation, two-input CNN.**

I. Introduction

Given a pair of face images, we ask the following questions: (i) are the faces distinguishable with respect to age, (ii) if so, what is the degree of distinguishability and (iii) which of the two faces is younger? We consider an image pair to be distinguishable if the ages differ by at least 2 years. By 'degree of distinguishability', we mean whether the difference in perceived ages between the two faces in a pair is large, medium, small, or tiny, where each of the attributes is defined by age differences in the ranges > 20, $(10, 20]$, $(5, 10]$ and $(2, 5]$, respectively. Fig. 1 shows examples of image pairs constructed for the purpose from the UTKFace [1] and Imdb-Wiki [2] datasets. Younger face identification is extremely critical to the cosmetics industry whose market is expected to reach over $\$800$ billion by 2023.

While the proposed problem is not addressed much in the literature, the closest related topic is that of facial age estimation [2]–[4]. Indeed, with recent progress in age estimation, the problem would seem to be easily addressed by applying an age estimation algorithm on each of the faces in the pair and deciding which is younger. However, as we show later, this late fusion [5] approach is prone to errors due to inaccuracies of the decision boundaries separating the large number of age categories (the mean absolute error of the state-of-the-art age estimation method [2] is 3.221). Ascertaining the degree of distinguishability is akin to classifying an image pair into the 4 aforementioned classes. In this task too, the errors make the decision fusion challenging. This kind of multi-class classification has also been addressed using ordinal regression [6] since there is an implied order in the 4 age differences. However, our approach does not require the generation of meaningful ordinal features, and we illustrate that softmax is more suited for our purpose.

Fig. 1: Example image pairs used for younger face recognition. From column 1 to 5 are image pairs indistinguishable, with large, medium, small and tiny age differences, the top row shows younger faces while the bottom row shows older faces.

Our method employs a two-stream convolutional neural network (CNN) that takes an image pair as input. The performance of multiple stream CNN could be strengthened if a particular stream has access to information from other streams. In some multi-input networks such as [7], [8], the streams get merged early in the network and there is no meaningful sharing of information among the streams, while in others [9], [10], the fusion is done at intermediate layers. Interaction between the two streams in the proposed network takes the form of a comparative layer that considers the mutual relationships of the feature vectors that are generated from the auxiliary classifiers connected to the inception layers of GoogLeNet [11]. We also construct a large face image pair dataset by permutating face images and corresponding age labels from the publicly available UTKFace [1] and Imdb-Wiki [2] datasets.

II. Related work

Early fusion schemes in multiple input CNNs involve stacking input images or optical flow maps [7]. Park et al. [12] present several late fusion schemes working on deep features and show best results for activity recognition by fusing the FC7 layers of multiple images based on VGG19 [9]. In the Siamese neural network (SNN) [5], contrastive loss is proposed to calculate the relationship between the two fully connected layers. Bell and Bala [13] minize the $L2$ distance of the fully connected layers of two parellel CNNs. Similarly, squared Euclidean distance is employed in [10]. Since order within image pairs is implicit in younger face identification, $L2$ distance based interactions are not applicable.

*Corresponding Author (ASDRajan@ntu.edu.sg)
Liangliang Wang and Deepu Rajan are with the Media & Interactive Computing Lab, School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798.

978-1-7281-3723-0/19/$31.00 ©2019 IEEE

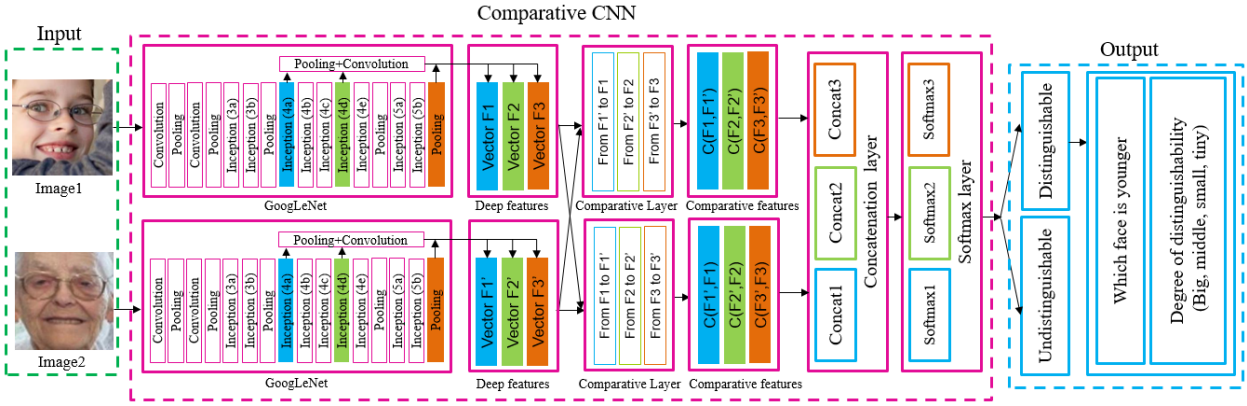


Fig. 2: Flowchart of proposed method.

In recent years, face age detection has been receiving attention, with approaches mainly falling into two categories: to estimate the actual age and the age range. In [14], Chao et al. propose to detect actual ages of images from FG-NET aging database [15] using a local regression scheme. On another face dataset named Adience [16], Levi and Hassner [3] introduce a new CNN architecture for age range detection.

III. Comparative CNN

A. Network Architecture

Fig. 2 shows the pipeline of the proposed algorithm as well as the architecture of the CNN used. The input to the pipeline is an image pair. Each of the two streams of the CNN is comprised of GoogLeNet [11]. Three feature vectors are constructed from each stream. The first is formed from the fully connected layer of the auxiliary classifier connected to the Inception (4a) layer followed by pooling and convolution, the second from a similar auxiliary classifier connected to the Inception (4d) layer and the third from final pooling layer of GoogLeNet. The deep features contained in the three vectors from the two streams interact correspondingly at the comparative layer to generate the comparative features, which are still three vectors from each stream. The comparative features are concatenated to form three vectors in the concatenation layer, each of which is passed through softmax regression. The output of the network consists of classification into distinguishable/indistinguishable and into the degree of distinguishability and also labeling the younger of the two faces in the image pair.

B. Modeling interaction between streams

As noted from the previous section, interaction between the two streams is manifested through the three feature vectors in each stream. For the sake of illustration, let us consider one vector from each stream - $F = [f_1, \cdots, f_n]$ and $F' = [f'_1, \cdots, f'_n]$ - with length $n = 1024$. Regarding F and F' as probably distributions over the values that the vector components can take, let X denote the discrete random variable and $p_1(x)$ and $p_2(x)$ denote the respective distributions of X. The KL divergence between the two distributions is given by $\sum_{x \in X} p_1(x) ln \frac{p_1(x)}{p_2(x)}$. However, in our formulation, we wish to capture the mutual relationship between each pair of corresponding vector components; in other words, we represent the relationship between the two vectors using pointwise distance. Thus, we consider $p_{1i} ln \frac{p_{1i}}{p_{2i}}$ where i denotes the i^{th} component. Next, in order to further strengthen the interaction between the two streams, we assign weights π_1 and π_2 to the two distributions such that π_1, $\pi_2 \geq 0$ and $\pi_1 + \pi_2 = 1$. The relationship can be rewritten as $\pi_1 p_{1i} ln \frac{\pi_1 p_{1i}}{\pi_2 p_{2i}}$. Clearly, this equation is undefined if $p_{2i} = 0$ and $p_{1i} \neq 0$, which means that the distribution $p_1(x)$ must be *absolutely continuous* [17] with respect to $p_2(x)$. This problem is overcome by defining the directed divergence [18] as

$$K(p_1, p_2) = \pi_1 p_1 \ln \frac{\pi_1 p_1(x)}{\pi_1(x) p_1(x) + \pi_2 p_2(x)}. \quad (1)$$

In our case, we write the corresponding pointwise expression as

$$K(p_{1i}, p_{2i}) = \pi_1 p_{1i} \ln \frac{\pi_1 p_{1i}}{\pi_1 p_{1i} + \pi_2 p_{2i}}. \quad (2)$$

If $\pi_1 = \pi_2 = 1/2$, then $K(p_{1i}, p_{2i}) = 0$ if and only if $p_{1i} = p_{2i}$, which satisfies the characteristics of a difference measure.

Each pointwise difference measure is collected into a comparative feature vector $C(F1, F1')$. The asymmetry of the KL divergence carries over to $C()$, which is also desirable since there is an ordinal relationship between the two images in the pair. Thus, a similar difference measure $C(F1', F1)$ is defined from $F1'$ to $F1$. Note that the Jensen-Shannon divergence can easily be derived from $C(F1, F1')$ and $C(F1', F1)$. Recall from Fig. 2 that 3 pairs of $C()$ feature vectors are formed and each corresponding pair is concatenated into a 2048 dimensional vector, which are input to the corresponding linear layers for softmax regression. The maximum of the three regressions is taken to be the output.

C. Training Methodology

Since our dataset is large enough and our problem is very distinct, we train the comparative CNN from scratch. To reduce the complexity of our architecutre as well as to compare

the two images on the same scale, the two parallel GoogLeNet architectures share the same parameters. The base learning rate is set as 0.001 and the batch size is set as 32 to balance the accuracy and memory cost. A total of $300,000$ iterations are implemented for the task of which face is younger, and $450,000$ iterations for the other two tasks. Caffe [19] is employed on a Nvidia Tesla v100 GPU.

IV. Experimental Results

A. Dataset Construction

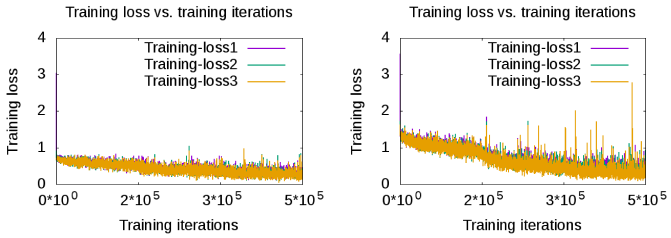
We collect face images from UTKFace [1] and Imdb-Wiki [2] datasets. From $23,708$ face images aged 1 to 116 in the UTKFace dataset, we choose $23,668$ images after removing non-face and wrongly annotated images. Faces more than 100 years old are reannotated as 100. To obtain a uniform distribution of ages in the combined dataset, we choose 50,456 Imdb-Wiki images after cropping them using a face detector [20]. This resulted in a total of $74,124$ images resized to 224×224.

We describe construction for the 'large' category in which the image pairs differ by at least 20 years. We uniformly sample images from 1 to 60 years to generate $184,206$ younger images in a pair. The older image of a pair is chosen by uniformly sampling images in the range $y + 21$ to 100 years, where y is the age of the younger image. A similar process is employed for age ranges $[1, 70]$ and $[1, 80]$ so that the other three categories are formed: $212,339$ image pairs in 'medium' $((10, 20])$, 240,467 image pairs in 'small' $((5, 10])$ and the same number in the 'tiny' $((2, 5])$ category. Thus, there are $877,479$ image pairs with at least 2 years difference and hence, distinguishable. We form the indistinguishable category where the age range is in $(0, 2]$, in which also there are $877,479$ images. Note that an image pair may contain the same subject.

B. Evaluation of Comparative CNN

From the collected $1,754,958$ face image pairs, we use one half as positive and the other half as negative samples for the task of identifying whether image pairs are distinguishable. The positive image pairs are also used for the task of identifying the degree of the distinguishability. To identify which face is younger, the distinguishable image pairs are further divided into two classes equally according to the sequence of young and old images within the image pair, where the positive image pair is defined as young image followed by old image while the negative is old followed by young. For every task, 90% samples are used for training and the rest for testing.

We first evaluate comparative CNN to determine if a face image pair is distinguishable. Three training losses corresponding to the classifiers in GoogleNet are shown in Fig. 3(a), which shows a gradual decrease with number of iterations since the dataset is large. The training losses for the task of determining the degree of distinguishability are shown in Fig. 3(b). Similarly, the training losses for the task of determining the younger face from each of the four categories is shown in Fig. 4. Convergence is faster when the age difference in a pair is large (Fig. 4(a)) and is slower for



(a) (b)

Fig. 3: Training evolutions of comparative CNN for recognizing (a) whether two faces are distinguishable and (b) what is the degree of distinguishability.

medium and small differences (Fig. 4(b) and (c)). However, comparative CNN does not converge for image pairs with tiny age difference suggesting that the task of determining younger face is challenging when the age difference is less than 5 years.

The performance for the three tasks are given in table 1. The proposed comparative CNN can correctly identify 71.63% of the image pairs as distinguishable or indistinguishable. We note that identifying indistinguishable image pairs is more challenging due to its small range of age difference. For identifying the younger face, the identification rate is proportional to the age difference, and the misclassification of image pairs with small and tiny age differences contribute most of the errors of determining the degree of distinguishability, whose average recognition rate is about 67%.

C. Comparison with other schemes

To reveal the advantage of comparative CNN, we compare it with four schemes: 1) stacking image pairs as input to a single GoogLeNet, 2) using optical flow color maps as input, 3) removing the comparative layer of comparative CNN, and directly concatenating the two fully-connected layers for softmax, 4) replacing the comparative and concatenation layer in comparative CNN with a subtraction layer for softmax. 5) Estimating the ages of two faces separately using a state-of-the-art age estimation approach [2], and then determining the younger one. We follow [21] for optical flow computation. From Table 1, we see that performance of age estimation algorithm of [2] in distinguishable/indistinguishable detection is unsatisfactory possibly because the 2 year age difference in an image pair is being poorly detected. In keeping with the trend of comparative CNN, all schemes improve performance in identifying younger face going from tiny to large differences.

Considering the task of determining the degree of distinguishability as an ordinal regression problem, we classify the deep features learned by the above schemes by replacing softmax with an ordinal cross entropy loss [6]. The highest recognition rate of 64.89% is obtained from using ordinal regression in the architecture that employs the subtraction layer. Since comparative CNN does not involve ordering of features, as expected, it does not perform well under an ordinal regression framework.

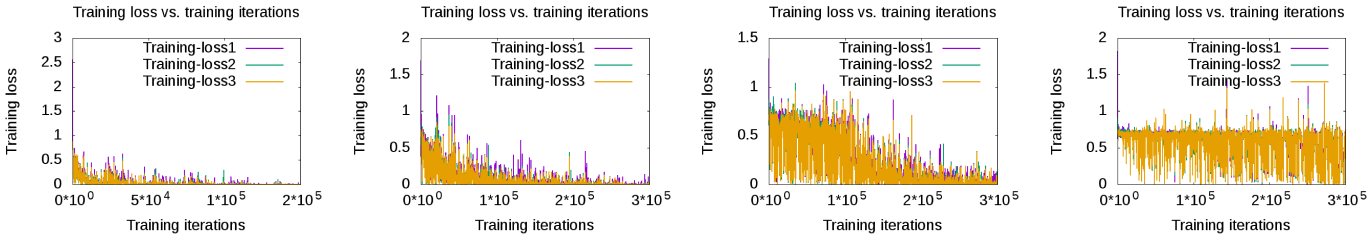


Fig. 4: Training evolutions of comparative CNN for recognizing which face is younger with (a) large, (b) medium, (c) small and (d) tiny age difference.

TABLE I: Average recognition rates of different CNN based methods.

Schemes	Whether distinguishable	What degree	Which face is younger			
			Large	Medium	Small	Tiny
Image stack as input [7]	63.68%	61.55%	90.65%	70.50%	56.69%	49.97%
Optical flow as input [7]	60.75%	58.00%	79.18%	67.65%	53.38%	49.92%
Remove comparative layer	69.37%	65.30%	96.85%	74.39%	60.86%	51.22%
Subtraction layer	69.97%	63.57%	96.77%	74.28%	59.38%	51.82%
Age estimation [2]	44.28%	32.97%	93.33%	77.63%	55.39%	48.24%
Comparative CNN	71.63%	66.77%	97.27%	75.72%	62.83%	53.05%

V. Conclusion

This paper presented a comparative CNN for the problem of whether a pair of face images is distinguishable in terms of age, if distinguishable, to what degree, and which face is younger. The comparative CNN is comprised of two parellel GoogLeNet architectures up to the last pooling layers, a comparative layer for computing the mutual relationships of the two deep features, a concatenation layer to combine the two comparative features, and a softmax layer for classification. We demonstrated our architecture on a new dataset that contains over 1.7 million image pairs with young/old labels.

VI. Acknowledgement

This research was funded by A*Star BMRC under the grant A*Star BMRC SPF Grant No. APG2013/057.

References

[1] Z. Zhang, Y. Song, and H. Qi, "Age progression/regression by conditional adversarial autoencoder," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 4352–4360.

[2] R. Rothe, R. Timofte, and L. V. Gool, "Deep expectation of real and apparent age from a single image without facial landmarks," *International Journal of Computer Vision*, vol. 126, no. 2-4, pp. 144–157, April 2018.

[3] G. Levi and T. Hassncer, "Age and gender classification using convolutional neural networks," in *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, June 2015, pp. 34–42.

[4] S. Chen, C. Zhang, M. Dong, J. Le, and M. Rao, "Using ranking-cnn for age estimation," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 742–751.

[5] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *2015 International Conference on Machine Learning (ICML)*, Jul 2015, vol. 37.

[6] Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua, "Ordinal regression with multiple output cnn for age estimation," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 4920–4928.

[7] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *2014 International Conference on Neural Information Processing Systems (NIPS)*, Oct 2014, pp. 568–576.

[8] Y. Sun, L. Zhu, G. Wang, and F. Zhao, "Multi-input convolutional neural network for flower grading," *Journal of Electrical and Computer Engineering*, vol. 2017, pp. 1–8, 2017.

[9] A Zisserman K Simonyan, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2014.

[10] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu, "Learning fine-grained image similarity with deep ranking," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, June 2014, pp. 1386–1393.

[11] C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 1–9.

[12] E. Park, X. Han, T. L. Berg, and A. C. Berg, "Combining multiple sources of knowledge in deep cnns for action recognition," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, March 2016, pp. 1–8.

[13] S. Bell and K. Bala, "Learning visual similarity for product design with convolutional neural networks," *ACM Transactions on Graphics*, vol. 34, no. 4, pp. 98:1–98:10, Aug 2015.

[14] Wei-Lun Chao, Jun-Zuo Liu, and Jian-Jiun Ding, "Facial age estimation based on label-sensitive learning and age-oriented regression," *Pattern Recognition*, vol. 46, no. 3, pp. 628 – 641, 2013.

[15] "The fg-net aging database," Available at http://www.fgnet.rsunit.com, Accessed Nov, 2014.

[16] E. Eidinger, R. Enbar, and T. Hassner, "Age and gender estimation of unfiltered faces," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 12, pp. 2170–2179, Dec 2014.

[17] S. Kullback, *Information Theory and Statistics*, Dover Publications, New York, 1968.

[18] J. Lin, "Divergence measures based on the shannon entropy," *IEEE Transactions on Information Theory*, vol. 37, no. 1, pp. 145–151, Jan 1991.

[19] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. B. Girshick, S. Guadarrama, and T. Darrell, "Caffe: convolutional architecture for fast feature embedding," in *ACM Multimedia*, 2014.

[20] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. (CVPR)*, Dec 2001, vol. 1, pp. 511–518.

[21] Thomas Brox and Jitendra Malik, "Large displacement optical flow: descriptor matching in variational motion estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, pp. 500–513, 03 2011.