

DEEP EMBEDDING NETWORK FOR ROBUST AGE ESTIMATION

Yating He^{1,2}, Min Huang^{1*}, Qinghai Miao¹, Haiyun Guo^{1,2} and Jinqiao Wang^{1,2}

¹University of Chinese Academy of Sciences, Beijing, China, 100049

²National Laboratory of Pattern Recognition, Institute of Automation
Chinese Academy of Sciences, Beijing, China, 100190

heyating15@mails.ucas.ac.cn, {huangm, miaoqh}@ucas.ac.cn, {haiyun.guo, jqwang}@nlpr.ia.ac.cn

ABSTRACT

Estimating age through a single facial image is a classic and challenging topic in computer vision. Since facial images of the same age vary considerably, while those from different ages may look very similar. To address these problems, we propose an end-to-end deep embedding neural network for robust age estimation. Specifically, we jointly use classification loss and triplet-based ranking loss to train a deep embedding network, which maps the input facial images into an embedding metric space where features of the same age are compact and those from different ages are pushed away. Thus the deep embedding network can learn more discriminative features and improves the performance for age estimation. Additionally, to accelerate the convergence of the network, we adopt an online hard negative mining strategy during the triplet loss computation. Experimental results on public datasets MORPH II and FG-NET show the superiority of our approach compared to the state-of-the-art.

Index Terms— age estimation, convolutional neural network, multi-task loss.

1. INTRODUCTION

Age estimation is to predict age from a single facial image. The contemporary society has witnessed the miraculous development of artificial intelligence, which provides many effective approaches for machine to estimate age more precisely. Automatic age estimation from facial images has a lot of potential applications, including demographic statistics collection, commercial user management and assistance of biometrics systems, etc.

At the beginning, researchers simply classified facial images into several age groups such as infants, young adults, and senior adults [1]. Nowadays, most studies focus on the precise age from 0 to 100. Age estimation is a challenging topic in computer vision. Estimating a person's age only by facial appearance is difficult not only for humans but also to machines. The main reason is facial appearance of different people of the same age vary considerably. There are numerous factors that affect the facial appearance. Firstly, different

people has different aging speeds. Besides gender is another factor that affects age estimation. For example, a woman may look younger or older than a man of the same age. Even for the same person, his facial appearance changes slower in some years but faster in other years. In most cases, the changes of appearance of a person within a year are minimal and it is difficult to tell if Jack is 40 years old or 41 from one picture. Moreover, the captured facial images are also affected by pose, illumination and occlusion, which increase the difficulty of age estimation.

There has been lots of works for age estimation. Kwon *et al.* [1] only classified input images into one of three age groups according to some hand-crafted features through skin wrinkle analysis and facial geometry features. Geng *et al.* [2, 3] proposed an Aging pattern Subspace (AGES) approach to define an images sequence of one subject as an aging pattern based on PCA model which obtained the mean absolute error (MAE) of 6.22 on FG-NET [4] database. Many nonlinear regression approaches, such as quadratic regression [5], Support Vector Regression (SVR) [6] and Gaussian Process [7, 8], have been used for age estimation. However, Chang *et al.* [9] supposed that learning non-stationary kernels for a regression problem could easily cause over-fitting. Recently, some deep learning methods have been proposed to improve the performance of age estimation. Rothe *et al.* [10] proposed a Deep EXpectation (DEX) model and trained CNN to classify input facial images into one of 101 classes. Although the learned CNN feature involves more semantic information compared to the above hand-crafted features, it's still not discriminative enough since the classification loss only focuses on pulling features from different ages apart and ignores the great variations between features of the same age.

In this paper, we propose an end-to-end deep embedding network for age estimation. Specifically, we utilize the great discriminative power of CNN to abstract efficient CNN features from facial images, which involve more semantic information than traditional hand-crafted features. Since different facial images of the same age vary considerably but those from different age may look very similar. To deal with the complex variations and learn an efficient metric space where features of the same age is closer than those

* corresponding author

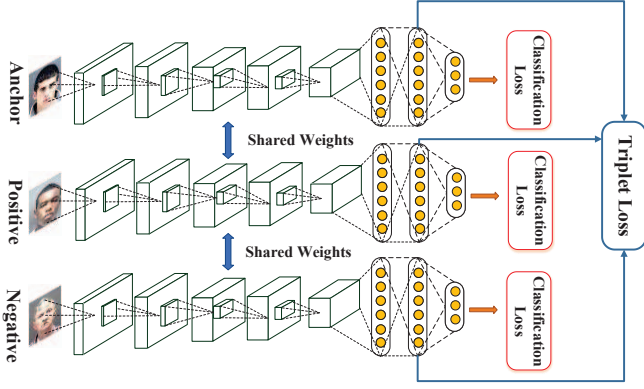


Fig. 1. The overall framework of deep embedding neural network for age estimation.

from different age, we adopt a multi-task learning framework and jointly optimize the CNN with classification loss and triplet loss. On the one hand, classification loss can effectively pull the CNN features of different ages apart. On the other hand, for each triplet pair, which consists of two facial images of the same age and one from a different age, triplet loss aims to separate the positive pair from the negative by a distance margin, not only enlarging the distance between the negative pair but also reducing the distance between the positive.

With the joint optimization of classification loss and triplet loss, deep embedding network can learn more discriminative classifier for age estimation. Besides, in the embedding metric space, the Euclidean distance of features directly corresponds to the age gap of facial images, depicting the ordinal relationship of different ages to some extent. Finally, an online hard negative mining strategy is utilized during the triplet loss computation to accelerate the convergence of deep embedding network.

2. DEEP EMBEDDING NETWORK

Since introduced by LeCun [13] in early 1990's, CNN has demonstrated the record-beating performance at challenging tasks such as image classification, object detection and face recognition. In this paper, we learn a nonlinear projection from the input facial image space to a facial age feature space using a deep embedding convolution network. Specifically, given two input facial images I_b and I_c , we calculate the similarity of them based on the squared Euclidean distance between the features of them in the embedding feature space as:

$$D(I_b, I_c) = \|f(I_b) - f(I_c)\|_2^2 \quad (1)$$

Given two images, the smaller the Euclidean distance $D(I_b, I_c)$ is, the more similar they are.

The overall framework of deep embedding neural network is illustrated in Fig.1. Given a triplet (X_t^a, X_t^p, X_t^n) , where X_t^a , X_t^p and X_t^n are the anchor, positive and negative image of the t -th triplet respectively. Positive has the same age as the anchor image, while the negative has a different age. Triplet loss aims to separate the positive pair of the triplet from the negative by a distance margin, which is defined as:

$$T(X_t^a, X_t^p, X_t^n, W_T) = \max\{0, m + D(X_t^a, X_t^p) - D(X_t^a, X_t^n)\} \quad (2)$$

where W_T is the weight parameter during the feature extraction process and m is the margin parameter. In our experiments, we set m to 0.5. Triplet loss can not only enlarge the distance between features from different ages but also reduce that of the same age.

Classification loss is defined as:

$$C(I, l; W_T, W_S) = -\log \hat{p}_l \quad (3)$$

where l is the groundtruth age label of input facial image I and W_S is the parameter weight between the feature extraction layer and the classification loss layer. \hat{p}_l is the predicted probability of image I belonging to age label l . Classification loss focuses on pulling the features of different ages apart. The joint use of classification loss and triplet loss can enable the CNN to learn a metric space, where the distance between features of the same age is smaller than that from different ages. Thus the learned CNN features are more discriminative for age estimation and the final classification performance will be boosted.

We adopt the basic VGG-16 architecture [14] for the deep embedding network. In the train phase, to train the deep embedding network efficiently, we adopt an effective online triplet selection scheme. Since there are many triplets satisfying the constraint defined in Eq.2, thus they have no contribution to the gradient update during the network training. Then their participation in the network forward and backward pass will slow down the network convergence. Therefore, we adopt an online hard negative mining strategy like [14] which uses all anchor-positive pairs in a mini-batch while selecting the hard negatives during the triplet loss computation.

More specifically, we choose all combinations of each anchor-positive pairs and all corresponding negatives as triplets and select the top K ones with the highest loss values. To reduce the amount of computational cost, we back-propagate the sum of losses on these top K triplets as the final triplet loss to update the network parameters.

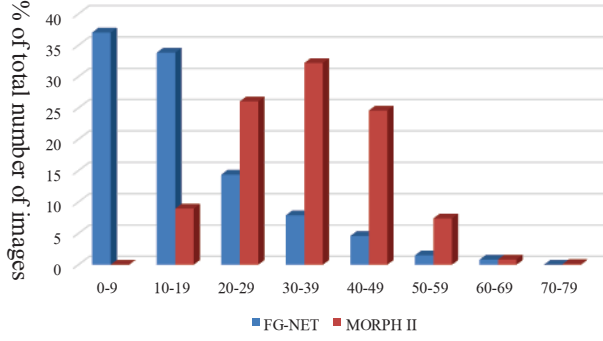


Fig. 2. Age range distribution of face image in the FG-NET and MORPH II database.

Finally, we combine the classification loss and triplet loss as a multitask loss in a linear weighted way. Our multi-task loss function can be represented as:

$$\min L(W_T, W_S) = \frac{\alpha}{2} \cdot \|W_T\|_F^2 + \frac{\alpha}{2} \cdot \|W_S\|_F^2 + \beta \sum_{m=1}^M C(I_m, I_m; W_T, W_S) + \sum_{t=1}^T T(X_t^a, X_t^p, X_t^n, W_T) \quad (4)$$

where α is the weight decay and β is hyper-parameter which can balance the role of each loss in the network. M denotes the number of images in a mini-batch and T is the number of selected triplet during online hard negative mining. In our experiments, we set α to 0.0005 and β to 0.4.

In the test phase, to estimate the real age more accurately, we compute a softmax expected value refinement [10], and the final predicted age is Y :

$$E(Y) = \sum_{i=0}^{100} P_i Y_i \quad (5)$$

where P_i is the predicted probability of the corresponding age Y_i , and $Y = \{0.1, \dots, 100\}$ represents 101 age categories.

3. EXPERIMENTS

3.1. Datasets

We evaluate the deep embedding neural network on two benchmark age databases: MORPH II [12] and FG-NET [4]. The former one is a publicly available dataset and widely used for age prediction. MORPH II contains 55,000 color facial images of 5,475 individuals whose age ranges from 16 to 77. FG-NET contains 1,002 colour or gray facial images of 82 individuals whose age are ranging from 0 to 69. Fig.2 shows the age range distribution of the facial images in the two databases.

3.2. Experimental setting

In our experiments, we follow the practice in work [9, 16, 17] and randomly sampled 80% of MORPH II as training set and the remaining as testing set. It is ensured that there is no overlapping between the training and testing images.

No matter in the training phase or in the testing phase, the input of the network is RGB images that are resized into 224×224 . We use stochastic gradient descent for network training and the mini-batch size is set to 64. The models are trained for 297000 iterations and the learning rate starts from 0.001 with the GTX TITAN X GPU.

In order to enhance the performance of the proposed method on MORPH II and FG-NET, our networks are pre-trained on IMDB-WIKI database [10] which contains 0.5 million images of celebrities from IMDB and Wikipedia. In our pre-training phase, all images in IMDB-WIKI database would be processed by an off-the-shelf face detector of Mathias *et al.* [18] to obtain the location of the face and non-face images would be removed. In our whole experiments, no face alignment techniques are used to detect the facial landmarks.

3.3. Experimental results and analysis

We measure the performance of age estimation by the Mean Absolute Error (MAE) that is calculated using the average of the absolute errors between estimation age and chronological age in our work. More specifically, the MAE can be calculated as following:

$$MAE = \frac{\sum_{i=1}^N |P_i - Q_i|}{N} \quad (6)$$

where P_i and Q_i are the estimation age and chronological age of the i -th image.

In the test phase, to estimate the real age more accurately, we compute a softmax expected value as our final predicted age. We conduct our experiment with the classification loss function and the multi-task loss function respectively and show the results in Table 1. Compared with the classification loss, the model with multi-task loss reduces the MAE dramatically from 2.78 to 2.71 and from 3.25 to 3.19 on MORPH II and FG-NET databases. It is evident that the performance of the multi-task loss function is superior to the classification loss under MAE.

Table 1. Results of our deep embedding neural network with different loss functions on the MORPH II and FG-NET datasets (the lower the better).

Methods	MORPH II	FG-NET
Classification loss	2.78	3.25
multi-task loss	2.71	3.19

MORPH II							
Real age	18	25	29	32	35	37	50
Predicted age	18.5	24.2	28.8	31.6	34.4	37.5	50.6

Fig. 3. Examples of facial images from MORPH II with age estimation by our deep embedding neural network.








FG-NET							
Real age	3	6	17	27	30	36	61
Predicted age	3.02	6.3	16.6	28.3	31.4	36	61.8

Fig. 4. Examples of facial images from FG-NET with age estimation by our deep embedding neural network.

In Fig.3 and Fig.4, we show some examples of age estimation of our approach in the MORPH II and FG-NET dataset. We also compare our approach with the other methods [3, 9, 19] on the FG-NET and MORPH II aging datasets, and the comparison results are shown in Table 2 and Table 3.

Table 2. MAEs of different algorithms on the FG-NET aging dataset.

Method	MAE
AGES [3]	6.22
RankBoost [21]	6.02
LAR [22]	5.64
Hierarchical Framework [23]	4.97
PLO [19]	4.82
OHRANK [9]	4.48
Deep embedding network (ours)	3.19

From Table 2 and Table 3, we can see that our approach achieves the best performance with the MAE of 3.19 and 2.71 on FG-NET and MORPH II databases, respectively. The proposed deep embedding network reduces more than 1.0 in MAE compared to OHRANK [9] and Hierarchical Framework [23] on the FG-NET aging dataset and decreases a lot compared to OR-CNN [16] and CA-SVR [24] on the MORPH II dataset. The proposed method achieves the state-of-the-art performance in the task of age estimation on both FG-NET and MORPH II databases.

Table 3. MAEs comparison of different algorithms on the MORPH II aging dataset.

Method	MAE
AGES [3]	8.83
MTWGP [8]	6.28
CA-SVR [24]	5.88
SVR [5]	5.77
OHRANK [9]	6.07
DLA [17]	4.77
OR-CNN [16]	3.27
Deep embedding network (ours)	2.71

4. CONCLUSIONS

In this paper, we have proposed an end-to-end deep embedding neural network for age estimation. We utilize the great power of CNN to abstract efficient features from facial images, which involve more semantic information. A multi-task loss is applied to learn more discriminative classifier for age estimation. An online hard negative mining strategy is utilized for accelerating training and further booting performance by computing more meaningful gradient. Experimental results on two public MORPH II and FG-NET dataset show the superiority of the proposed approach.

5. ACKNOWLEDGMENT

This work was supported by National Natural Science Foundation of China 61303155, and Project of Y65201JY00-18.

6. REFERENCES

- [1] Y. H. Kwon, "Age classification from facial images," in *CVPR.IEEE*, 1994, pp. 762-767.
- [2] X. Geng, Z.-H. Zhou, Y. Zhang, G. Li and H. Dai, "Learning from facial aging patterns for automatic age estimation," in *Proceedings of the 14th ACM international conference on Multimedia.ACM*, 2006, pp. 307-316.
- [3] 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*, vol. 29, no. 12, 2007, pp. 2234-2240.
- [4] A. Lanitis and T. Cootes, "FG-NET Aging Data Base," *Cyprus College*, 2002.
- [5] 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 Transactions on Image Processing*, vol. 17, no. 7, 2008, pp. 1178-1188.
- [6] G. Guo, G. Mu, Y. Fu and T. S. Huang, "Human age estimation using bio-inspired features," in *CVPR.IEEE*, 2009, pp. 112-119.
- [7] B. Xiao, X. Yang, H. Zha, Y. Xu and T. S. Huang, "Metric learning for regression problems and human age estimation," in *Pacific-Rim Conference on Multimedia*, Springer, 2009, pp. 88-99.
- [8] Y. Zhang and D.-Y. Yeung, "Multi-task warped gaussian process for personalized age estimation," in *CVPR.IEEE*, 2010, pp. 2622-2629.
- [9] K.-Y. Chang, C.-S. Chen and Y.-P. Hung, "Ordinal hyperplanes ranker with cost sensitivities for age estimation," in *CVPR.IEEE*, 2011, pp. 585-592.
- [10] R. Rothe, R. Timofte and L. Van Gool, "Dex: Deep expectation of apparent age from a single image," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2015, pp. 10-15.
- [11] A. W. Rawls and K. Ricanek Jr, "MORPH: Development and optimization of a longitudinal age progression database," in *European Workshop on Biometrics and Identity Management*, Springer, 2009, pp. 17-24.
- [12] Y. LeCun, L. Bottou, Y. Bengio and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, 1998, pp. 2278-2324.
- [13] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
- [14] X. Wang and A. Gupta, "Unsupervised learning of visual representations using videos," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2794-2802.
- [15] Z. Niu, M. Zhou, L. Wang, X. Gao and G. Hua, "Ordinal regression with multiple output cnn for age estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4920-4928.
- [16] X. Wang, R. Guo and C. Kambhamettu, "Deeply-learned feature for age estimation," in *WACV.IEEE*, 2015, pp. 534-541.
- [17] M. Mathias, R. Benenson, M. Pedersoli and L. Van Gool, "Face detection without bells and whistles," in *European Conference on Computer Vision*, Springer, 2014, pp. 720-735.
- [18] C. Li, Q. Liu, J. Liu and H. Lu, "Learning ordinal discriminative features for age estimation," in *CVPR.IEEE*, 2012, pp. 2570-2577.
- [19] P. Yang, L. Zhong and D. Metaxas, "Ranking model for facial age estimation," in *Pattern Recognition (ICPR), 2010 20th International Conference on*, IEEE, 2010, pp. 3404-3407.
- [20] K. Ricanek, Y. Wang, C. Chen and S. J. Simmons, "Generalized multi-ethnic face age-estimation," in *BTAS.IEEE*, 2009, pp. 1-6.
- [21] Y. Liang, X. Wang, L. Zhang and Z. Wang, "A hierarchical framework for facial age estimation," *Mathematical Problems in Engineering*, vol. 2014, 2014.
- [22] 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*, 2013, pp. 2467-2474.