Deeply Learned Rich Coding for Cross-Dataset Facial Age Estimation

Zhanghui Kuang
SenseTime Group Limited
jeffrey.kuang@sensetime.com

Chen Huang
The Chinese University of Hong Kong
chuang@ie.cuhk.edu.hk

Wei Zhang
SenseTime Group Limited
wayne.zhang@sensetime.com

Abstract

We propose a method for leveraging publicly available labeled facial age datasets to estimate age from unconstrained face images at the ChaLearn Looking at People (LAP) challenge 2015 [9]. We first learn discriminative age related representation on multiple publicly available age datasets using deep Convolutional Neural Networks (CNN). Training CNN is supervised by rich binary codes, and thus modeled as a multi-label classification problem. The codes represent different age group partitions at multiple granularities, and also gender information. We then train a regressor from deep representation to age on the small training dataset provided by LAP organizer by fusing random forest and quadratic regression with local adjustment. Finally, we evaluate the proposed method on the provided testing data. It obtains the performance of 0.287, and ranks the 3^{rd} place in the challenge. The experimental results demonstrate that the proposed deep representation is insensitive to cross-dataset bias, and thus generalizable to new datasets collected from other sources.

1. Introduction

Automatic age estimation is an important topic in computer vision and multimedia. It conveys valuable facial information for age-specific advertising, vision-based demographics and many other applications. However, age estimation from face images captured in the wild still remains a challenging task due to the large variations of factors such as illuminations and poses. These factors, mixed with other difficulties in constrained facial age estimation such as gender, race and personal life style, further make the problem challenging.

Deep models have significantly boosted unconstrained face recognition in recent years [25, 22, 23, 20]. The great success of deep models is achieved by better deep network architectures and supervisory methods, as well as much larger training dataset. *e.g.*, DeepFace [25] was trained with about 4.4 million images, DeepID2+ [23] was trained with about 290 thousand images, and FaceNet [20] was trained with about 200 million images. However, the progress in unconstrained facial age estimation is much slower, due to the difficulty of collecting and labeling a large dataset. Several publicly available age datasets, including MORPH [15], FG-NET [1], and FACES [7], have been collected with real age information in control environment for academic research. Recently, Adience dataset [8] was constructed from Flickr photos with manually labeled age groups. In this paper, we propose a method which can leverage these existing facial age datasets to estimate age on a new dataset. We design novel supervisory signals which are insensitive to cross-dataset bias, and our deep representation can be transferred to the new dataset. The method is applied to ChaLearn Age Estimation Challenge organized at the International Conference on Computer Vision (ICCV 2015) [9]. The challenge organizer invented a novel application for the collaborative harvesting and labeling by the community in a gamified fashion, and finally collected about 5000 images with apparent age labeled by humans. Our method utilizes the power of deep models and avoids the small training set problem of this challenge.

Most of existing age estimation methods use handcrafted facial features, such as Local Binary Patterns (LBP) [2], Haar-like wavelets [31], and Active Appearance Model (AAM) [6] based feature. However, the handcrafted features, not specially designed for facial age estimation, only have low level information, and are lack of middle level and high level semantic meaning. Recently, some advanced features [14, 19] were also proposed. These features are usually projected to a low dimensional space by dimension reduction methods such as PCA, LDA, and manifold learning [10, 11], to construct more compact and discriminative features. These approaches generate the age-

related feature with a shallow model, and usually obtain sub-optimal results.

Herein, we propose a deep Convolutional Neural Network (CNN) based system to learn discriminative deep representation directly from image for age estimation. Its learning is supervised by rich binary codes, and thus modeled as a multi-label classification problem. We design our binary codes according to partitioning age into different groups at multiple granularities, and gender information. From deep representation to age, we learn a regressor by fusing Random Forest (RF) and Quadratic Regression with Local Adjustment (QRLA).

There exist a few deep models for age estimation in the literature [16, 29]. Kong *et al*. [16] learned mid-level features in an unsupervised manner. Yi *et al*. [29] trained CNN from local aligned face patches to simultaneously perform age estimation, gender prediction and ethnicity classification. Multi-scale analysis and data augmentation were applied for further improvement. Yang *et al*. [28] extracted features through a scattering network, then reduced the feature dimension by PCA, and finally predicted the age via category-wise rankers. Different from previous work, we utilize a very deep VGG network [21], and learn from the whole face images rather than patches. More importantly, we propose rich binary codes to guide the training process. Our rich codes not only characterize multi-source information of gender and age groups at several granularities, but also alleviate the problem caused by the labeling bias of multiple training datasets. Our experimental results demonstrate that our learned deep representation is robust against unconstrained facial variations, and also generalizable across different datasets.

The rest of this paper is organized as follows. Section 2 gives an overview of related work. Section 3 details the proposed method. Section 4 summarizes our experimental results, with concluding remarks in Section 5.

2. Related Work

In the literature, most of existing age estimation work can be categorized into two groups: the first ones focus on extracting good features while the second ones designing good predictors for age estimation.

Traditional age estimation methods rely on hand-crafted feature representation. Early studies [17] extract simple features of wrinkles and distances between facial components like eyes. LBP [2] and Haar-like wavelets [31] are also widely used. To extract more details, the popular AAM [6] learns both shape and appearance models by PCA. The state of the art Bio-inspired Features (BIF) [14] are constructed by convolving face images with Gabor filters of different scales and orientations followed by pooling. Scattering Transform (ST) [3] generalizes BIF by recovering the lost texture details in pooling. Recent deep CNN based work [16, 29] automatically learn powerful feature representation with classification or linear regression loss layers, and obtain impressive results. Different from previous work, our CNN model is supervised by rich binary codes during training.

When it comes to the prediction module, options include multi-way classification [14, 19, 26], regression [10, 11, 12, 13, 18, 30], and ranking [3, 4, 27]. Standard classification approaches, such as kNN [26], SVM [14] and RealAdaBoost [19] can be efficiently employed. However, classification methods greatly overlook the ordinal relationship between age values, but instead treat them as independent labels. Regression methods take this issue into account by learning a function that maps the feature space to the linearly increasing age-value space. Examples are Quadratic Model (QM) [10], Support Vector Regression (SVR) with local age adjustment using SVM [11] , Canonical Correlation Analysis (CCA) [13], Partial Least Squares (PLS) [12, 13], Gaussian Process (GP) [30] *etc*. Recently, ranking models achieved higher estimation accuracy by utilizing relative ordering information. The implicit assumption is that modeling linearly increased age values cannot reflect the non-stationarity of human aging process, and it is more appropriate to do a couple of relative comparisons of age pairs. Ordinal Hyperplanes Ranker (OHRank) [3, 4] as a representative, learns many binary SVM classifiers as rankers. Although the results are promising, the running speed is relatively slow. In this paper, we fuse two simple regressors for robustness, and more rely on our deeply-learned features to obtain good performance.

3. The Proposed Method

In this section, we describe our method for age estimation. The proposed deep CNN model is first introduced to learn age-related representation. The regressor from deep representation to age is then discussed.

3.1. Learning Deep Representation with Rich Codes

In the past several years, deep convolutional neural networks have demonstrated the ability to learn rich representation of scenes and objects. To make full use of the recent advances on visual representation learning, we employ a deep convolutional neural network (named as DeepCodeAge for simplicity) to learn age-related representation. In most of previous work, classification labels [21, 24] or regression real values [29] are used as supervisory signals to train deep models. Instead, we design 259-dimensional binary codes to train DeepCodeAge. The general idea is to encode different age groups at multiple granularities. We exhaustively test different partitions of age groups on the validation set and finally choose the set of partition configurations with the best performance. The first part is a 100-dimensional binary vector. Following the spirit of the

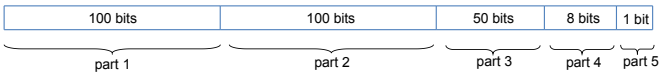


Figure 1. The layout of the binary vector used to train DeepCodeAge.

cumulative attribute space proposed in [5], its jth element a_j ($j \in \{1, 2, \cdots, 100\}$) for a face image with age y is set to 1 if $j \leq y$, otherwise 0. The second part is also a 100-dimensional binary vector. It is defined as the 1-of-k coding for multi-class classification with each age (from 1 to 100) being one category. Similarity, the third part is a 50-dimensional 1-of-k coding for classification with each consecutive two ages being one category. The fourth part is a 8-dimensional 1-of-k coding with 8 categories as defined in the Adience dataset [8]. Each element of the four binary vectors partitions the age line from 1 to 100 into two distinct parts. Those in the second, the third and the fourth vectors do so in the same way but at different granularities. Our experimental results validate that the DeepCodeAge trained with all the four type codes predicts age more accurately than that with only the first one. This probably is because the second, the third and the fourth type codes enforce two consecutive ages to be separable. The final part is a binary code indicating the gender. The underlying motivation is that gender classification and age estimation is not independent problems and doing them jointly can help each other as found in [29]. The whole binary vector is illustrated in Figure 1.

Network Architecture. Increasing the depth of deep models is essential for designing state of the art neural networks. Two representative very deep convolution neural networks VGGNet [21] and GoogleNet [24] both achieve very good performance in visual recognition and win the ImageNet competition. We also use very deep architecture. Our DeepCodeAge employs the same convolutional layers as VGGNet. However, it differs in two ways. First, deepCodeAge only has two fully-connected layers instead of three. We find that DeepCodeAge with two full-connected layers converges much faster than that with three ones in experiments. Second, since one 259-dimensional binary vector may has multiple positive bits, softmax loss layer doesn't work any more. We use a cross entry loss layer instead of the 1000-way softmax loss layer in VGGNet. Each input of the cross entry loss layer is the response of one class in classification, and thus learning deep representation for age estimation can be cast as a multi-label classification problem. The architecture of DeepCodeAge is shown in Figure 2. In this work, we also explore several VGGNet variants and combine them to seek improvements. We may extend our work to combine networks with different styles, such as VGGNet and GoogleNet.

Training procedure of DeepCodeAge. We train DeepCodeAge based on the pre-trained VGGNet model [21]. The VGGNet model was initially trained for general object recognition, which is much different from the facial analysis task we targeted at. We propose a two stage training procedure to train DeepCodeAge. In the first stage, we finetune the DeepCodeAge from the pre-trained VGGNet model to make it descriptive of facial features. We achieve this by training DeepCodeAge to recognize face on CelebFaces+ dataset [23] by classifying each face image into 1 of n ($n \approx 10000$) classes. In this stage, we use n-way softmax instead of the 259-way cross entropy to predict the probability distribution over n different identities. In the second stage, we train DeepCodeAge with the 259-dimensional binary vector based on the model learned in the first stage.

3.2. Regressor from Deeply Learned Representation to Age

The proposed DeepCodeAge can extract very rich representation for age estimation. Summarizing cumulative attribute space based codes is a straight-forward way to obtain the age. In our experiment, we find that a data-driven regression approach is better to explore the rich codes, and thus produce a more accurate estimation. We learn a regressor from deep representation to age by fusing two regressors. The first one is Quadratic Regression with Local Adjustment (QRLA) as done in [11]. The second is Random Forest (RF). Note that the above two regressors are trained only on the dataset provided by the organizer in the Challenge. The predict results of the two regressors are averaged as the final prediction.

4. Experiments

In this section, we first describe the dataset provided by the organizer at the ChaLearn LAP challenge 2015, those we used to train DeepCodeAge, and evaluation protocol. Then we give a detailed description of the implementation details about face detection, face normalization, and training procedure. Finally, we present and analyze the experimental results of the proposed method on the dataset of the challenge.

4.1. Datasets and Evaluation Protocol

The dataset provided by the organizer of the challenge is composed of 2476 training images, 1136 validation images and 1079 testing images. These face images are captured in different conditions in terms of viewpoints, image resolutions, and facial expressions. Each image is manually labeled by at least 10 people, and its average labeled age and standard deviation is computed.

Although the above dataset is valuable, it has very limited number of images, and training the proposed DeepCodeAge on it only easily leads to overfitting. To this end, during the training procedure of DeepCodeAge, we use a set

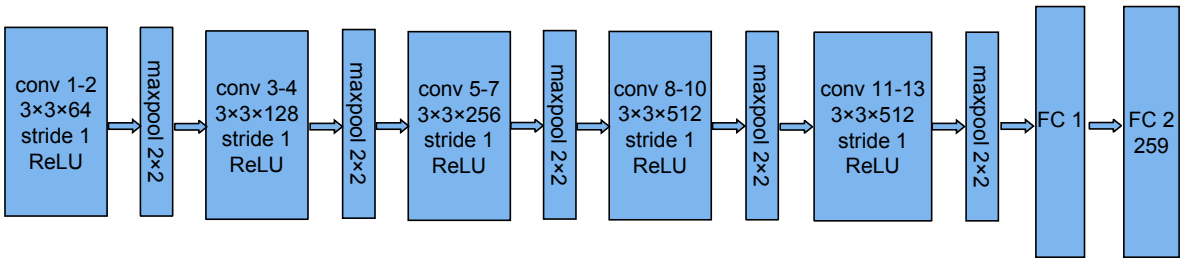


Figure 2. The architecture of our deep CNN for age estimation.

of academic datasets including MORPH [15], FG-NET [1], Adience [8], and FACES [7].

For one image with the average labeled age μ and the standard deviation δ, its evaluation criterion is computed by

$$\epsilon = 1 - e^{-\frac{(x-\mu)^2}{2\delta^2}}, \quad (1)$$

where x is the predicted age. The performance value of a method on a dataset is the average value of the evaluation criteria of its all images. From Equation 1, the smaller the performance value is, the better the accuracy of age estimation is.

We also use the mean of absolute error (MAE) to evaluate the performance of the proposed method.

4.2. Implementation Details

We detect faces and 21 facial landmarks using a commercial face SDK. In order to increase the recall of face detection, we detect faces and facial landmarks on flipped and rotated images, and select the one with highest confidence and a relatively large size if multiple faces are detected. In this way, we can automatically detect one face for all images provided by organizer without any manual effort.

We normalize each face to a canonical face with the same approach in [22], and then crop a $m \times m$ face region and resize it to a 224×224 image which is fed into DeepCodeAge.

During the training procedure of DeepCodeAge, we augment input images by horizontal flipping, converting RGB images to gray images (and then replicating a gray image three times so that the augmented image has 3 channels), and randomly translating the cropped region. By setting the number of hidden neurons of the FC 1 layer of DeepCodeAge to 384 and 4096, we use two variants of DeepCodeAge called DeepCodeAge$_{384}$ and DeepCodeAge$_{4096}$ respectively.

During training procedure of the regression model, for each face image, we concatenate the deep representation generated by DeepCodeAge$_{384}$ and DeepCodeAge$_{4096}$ on its RGB image and its corresponding gray image with two cropping sizes, resulting in a 2072-dimensional ($259 \times 2 \times 2 \times 2$) feature vector. These feature vectors are projected into a 50-dimensional space by PCA for regressor training. For each image, we randomly translate the cropped region 20 times for each cropping size. Therefore, the total number of training samples for regressor increases greatly.

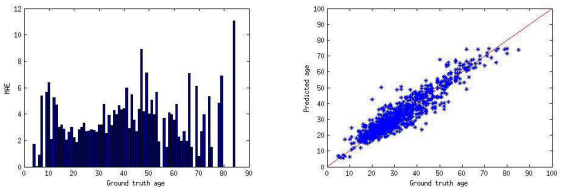
4.3. Experimental Results

We first evaluate the effects of different architecture fusion, and regressor fusion on the performance of age estimation. In these experiments, the regressor from deep representation to age is trained on the training set and the performance is evaluated on the validation set. We then evaluate the whole method on the testing set. In this experiment, the regressor is trained on both the training set and the validation set.

Evaluation of fusing different architectures. We investigate the performance of DeepCodeAge with different architectures. For simplicity, in this experiment, only RF is used as the regressor. we design two settings: (1) only the representation of DeepCodeAge$_{4096}$ is used to train the regressor; (2) the concatenation of representation of DeepCodeAge$_{4096}$ and DeepCodeAge$_{384}$ is used to train the regressor. The proposed method obtains the score of 0.296 in the first setting while 0.289 in the second setting. From experimental results, it is clear that fusing different architectures can boost the performance of age estimation.

Effectiveness of fusing RF and QRLA. On the validation set, our method with RF as the regressor achieves the score of 0.289 and the MAE of 3.32 while the one with QRLA achieves the score of 0.291 and the MAE of 3.32. The simple fusion by averaging the predictions of RF and QRLA achieves the score of 0.285 and the MAE of 3.29, which is slightly better than of the proposed method with a single regressor. Figure 3 shows the MAE for each age, and predicted age against the ground true age on the validation set using the fused regressor. We have two observations: (1) the ages of most of face images can be predicted within a small error; (2) the ages of old people and children are predicted with bigger errors than those of middle-aged people. This is probably due to very few training samples available for old people and children in the challenge.

Evaluation results on the testing set of the Challenge. Based on numerical evaluation and analysis above, we conclude that both fusing different architectures and different regressors contribute to performance improvement. Therefore, our challenge solution employs the above two strate-



(a) (b)

Figure 3. The performance of the proposed method on the validation set. (a) shows the MAE for each age. (b) shows the predicted age against the ground true age.

Table 1. Comparison the performance of the proposed method with that of other teams.

Rank	Team	Score
1	CVL_ETHZ	0.265
2	ICT-VIPL	0.271
3	**AgeSeer**	0.287
4	WVU_CVL	0.295
5	SEU-NJU	0.306

gies. The challenge results are shown in Table 1. Our method ranks the 3^{rd} place.

5. Conclusions

This paper has presented an effective method for age estimation from face images captured in the wild. We have designed rich binary codes which encode age groups at multiple granularities and gender information to learn deep representation for age estimation. We have shown that the learned deep representation work well on a new data set which is collected from other sources. In the future, we would like to investigate other effective supervisory signals for deep representation learning of age estimation.

References

[1] FG-NET (Face and Gesture Recognition Network) www.prima.inrialpes.fr/FGnet. 1, 4

[2] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *TPAMI*, 28(12):2037–2041, 2006. 1, 2

[3] K.-Y. Chang and C.-S. Chen. A learning framework for age rank estimation based on face images with scattering transform. *TIP*, 24(3):785–798, 2015. 2

[4] K.-Y. Chang, C.-S. Chen, and Y.-P. Hung. Ordinal hyperplanes ranker with cost sensitivities for age estimation. In *CVPR*, 2011. 2

[5] K. Chen, S. Gong, T. Xiang, and C. C. Loy. Cumulative Attribute Space for Age and Crowd Density Estimation. In *CVPR*, number 2, pages 2467–2474, 2013. 3

[6] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. *TPAMI*, 23(6):681–685, 2001. 1, 2

[7] N. Ebner, M. Riediger, and U. Lindenberger. Faces - a database of facial expressions in young, middle-aged, and older women and men: Development and validation. In *Behavior Research Methods*, pages 351–362, 2010. 1, 4

[8] E. Eidinger, R. Enbar, and T. Hassner. Age and Gender Estimation of Unfiltered Faces. *IEEE Transactions on Information Forensics And Security*, pages 1–10, 2013. 1, 3, 4

[9] S. Escalera, J. Fabian, P. Pardo, X. Baró, J. Gonzàlez, H. J. Escalante, Isabelle, and Guyon. Chalearn 2015 apparent age and cultural event recognition: Datasets and results. In *ICCV ChaLearn Looking at People workshop*, 2015. 1

[10] Y. Fu and T. Huang. Human age estimation with regression on discriminative aging manifold. *TMM*, 10(4):578–584, 2008. 1, 2

[11] G. Guo, Y. Fu, T. Huang, and C. Dyer. Locally adjusted robust regression for human age estimation. In *WACV*, 2008. 1, 2, 3

[12] G. Guo and G. Mu. Simultaneous dimensionality reduction and human age estimation via kernel partial least squares regression. In *CVPR*, 2011. 2

[13] G. Guo and G. Mu. Joint estimation of age, gender and ethnicity: CCA vs. PLS. In *FG*, 2013. 2

[14] G. Guo, G. Mu, Y. Fu, and T. Huang. Human age estimation using bio-inspired features. In *CVPR*, 2009. 1, 2

[15] K. R. Jr. and T. Tesafaye. MORPH: A longitudinal image database of normal adult age-progression. In *International Conference on Automatic Face and Gesture Recognition*, pages 341–345, 2006. 1, 4

[16] S. Kong, Z. Jiang, and Q. Yang. Learning mid-level features and modeling neuron selectivity for image classification. *arXiv preprint*, arXiv:1401.5535, 2014. 2

[17] Y. H. Kwon and N. d. V. Lobo. Age classification from facial images. *CVIU*, 74(1):1–21, 1999. 2

[18] L. Li and H. Lin. Ordinal regression by extended binary classification. In *NIPS*, 2006. 2

[19] H. Ren and Z.-N. Li. Age estimation based on complexity-aware features. In *ACCV*, 2014. 1, 2

[20] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015. 1

[21] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 2, 3

[22] Y. Sun, X. Wang, and X. Tang. Deep Learning Face Representation by Joint Identification-Verification. In *NIPS*, pages 1–9, 2014. 1, 4

[23] Y. Sun, X. Wang, and X. Tang. Deeply learned face representations are sparse, selective, and robust. In *CVPR*, 2015. 1, 3

[24] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, 2015. 2, 3

[25] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*, 2014. 1

[26] B. Xiao, X. Yang, Y. Xu, and H. Zha. Learning distance metric for regression by semidefinite programming with application to human age estimation. In *ACM MM*, 2009. 2

[27] S. Yan, H. Wang, T. Huang, Q. Yang, and X. Tang. Ranking with uncertain labels. In *ICME*, 2007. 2

[28] H.-F. Yang, B.-Y. Lin, K.-Y. Chang, and C.-S. Chen. Automatic age estimation from face images via deep ranking. In *BMVC*, pages 55.1–55.11, 2015. 2

[29] D. Yi, Z. Lei, and S. Z. Li. Age estimation by multi-scale convolutional network. In *ACCV*, 2014. 2, 3

[30] Y. Zhang and D.-Y. Yeung. Multi-task warped gaussian process for personalized age estimation. In *CVPR*, 2010. 2

[31] S. Zhou, B. Georgescu, X. S. Zhou, and D. Comaniciu. Image based regression using boosting method. In *ICCV*, 2005. 1, 2