Age Estimation by Multi-scale Convolutional Network

Dong Yi, Zhen Lei, and Stan Z. Li

Center for Biometrics and Security Research & National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences

Abstract. In the last five years, biologically inspired features (BIF) always held the state-of-the-art results for human age estimation from face images. Recently, researchers mainly put their focuses on the regression step after feature extraction, such as support vector regression (SVR), partial least squares (PLS), canonical correlation analysis (CCA) and so on. In this paper, we apply convolutional neural network (CNN) to the age estimation problem, which leads to a fully learned end-to-end system can estimate age from image pixels directly. Compared with BIF, the proposed method has deeper structure and the parameters are learned instead of hand-crafted. The multi-scale analysis strategy is also introduced from traditional methods to the CNN, which improves the performance significantly. Furthermore, we train an efficient network in a multi-task way which can do age estimation, gender classification and ethnicity classification well simultaneously. The experiments on MORPH Album 2 illustrate the superiorities of the proposed multi-scale CNN over other state-of-the-art methods.

1 Introduction

Human age estimation from face images is a young but hot research topic. The earliest work about age estimation was published in 1994 by Kwon and Lobo [1], in which the age was just classified into several ranges. After 2000, pushed by new models, features and classifiers in facial analysis, the field of age estimation started to flourish and the mean absolute error (MAE) on FG-NET database [2] was reduced from 9 to 5 years gradually. With moderate accuracy, age estimation can appeal the basic requirements of many applications, such as demographics analysis, commercial user management, and video security surveillance. Recently, Guo and Mu [3] first obtained a MAE below 4 years on the MORPH database [4]. This paper reduces the MAE further, which will supply a more stable age estimator for practical applications.

Like other facial analysis technics, age estimation is easily affected by many intrinsic and extrinsic factors. The most important factors include identity, gender and ethnicity. The face images of difference persons have different statistical properties, *i.e.,* For two persons of same age, maybe one has a young face, but another has an old face. The relationship between image pixels and age may be different when the face images come from different gender or ethnicity. Therefore,

the age of face image is tightly coupled with its identity, gender and ethnicity. Other factors are similar to those in face recognition, such as pose, illumination and expression (PIE). Due to the above reasons, age estimation is a hard problem and the relation between face image to age is highly nonlinear. It's hard to find a robust and accurate function to map the image pixels to its corresponding age.

Most existing methods estimate the age of face image by two steps: local feature extraction and regression (or classification). The task of local feature extraction is to get a representation robust to irrelevant factors list above, such as identity, gender, ethnicity, PIE and so on. And the dimension of the local features are usually reduced by feature selection or down-sampling. Based on the low dimensional features, regression methods are used to predict the age of face image, such as SVR [5], PLS [6] and CCA [7]. In this framework, the most representative work is BIF+CCA (or rCCA, KCCA) [3], which includes three steps: Gabor filters [8], Max+Std pooling and CCA. With careful tuning, this method achieved very high performance, but we still have room to improve the performance further.

In this paper, we propose a novel age estimation method based on convolutional neural network (CNN). Compared with BIF+CCA, CNN has learnable parameters and there is no gap between the feature extraction step and the regression step. In CNN, all steps are optimized together to minimize the estimation error. Note that Yang *et al.* [9] has used CNN for age estimation under surveillance scenarios, but the focus of their work is face tracking. For the age estimation module in the system, they just use the original CNN without much modification, therefore the accuracy of [9] is lower than BIF [10].

To dig out the power of CNN, we incorporate the tricks in traditional facial analysis methods into it. 1) Several facial landmarks are used to generate many local aligned patches from a face image and feed them into CNN, which can make our method more robust to image translation and pose variations. 2) Face image is cropped into many multi-scale patches and a regression function is learned on these patches jointly. Due to the complementary information between different parts and scales, multi-scale analysis can improve the performance of CNN significantly. 3) Facial symmetry is used to augment the database, which improves the generalization of CNN. Finally, we train a multi-task CNN to illustrate the flexibility of the proposed network. The multi-task network can estimate the age, gender and ethnicity of face image simultaneously in high precision and speed.

Because the complexity of CNN is higher than traditional methods, large data are needed to train a good network. In existing age databases, MORPH Album 2 has the largest scale, containing more than 55,000 face images. To alleviate the over-fitting problem of CNN, we choose MORPH Album 2 as the database for experiments. On this database, we achieve a new state-of-the-art result: MAE= 3.63 years, which is better than 3.98 years of BIF+KCCA [3]. Meanwhile, the speed of CNN is much faster than KCCA because of its big kernel matrix.

2 Related Works

Human age estimation from face image has been studied for 20 years. Limited by the technology of facial analysis, early methods mainly used geometric features to judge the age range of face image, such as baby, young adult and senior adult. Popular geometric features included chin drop, nose drop and so on [1],[11]. Geometry features can discriminate baby and adult easily but cannot distinguish adult and old man. Therefore, geometric and texture features were combined in some works. As the improvement of classification accuracy, researchers started to estimate the exact age instead of the coarse age range. Because AAM [12] was a natural tool to model the shape and texture of face image, many novel methods were proposed based on it, such as AAM+Quadratic Estimator[2], Aging Pattern Subspace (AGES) [13] and so on. By combing AGES and LDA [13], the MAE on FG-NET achieved 6.22 years. However, AAM is an pixel based method, which causes the AAM based methods unstable to environmental variations. After 2007, local features gradually became the mainstream in this field, such Gabor [14], LBP [15], Spatially Flexible Patch (SFP) [16], and BIF [10].

Based on these features, much attention were paid on the second step: age estimation by classification or regression. From the extracted features, we need predict the age range or the exact age from them. For classification, SVM and SVR are the most popular methods. Using BIF+SVM, [10] achieved MAEs of 3.47 and 3.91 years for male and female on YGA database [17]. In the same paper, the authors reported the MAE of 4.77 years on FG-NET by using BIF+SVR. In [18], Cao *et al.* formulated the age estimation as a ranking problem and proposed a novel method based on Rank-SVM [19]. On a subset of MORHP Album 2, they achieved a good result, MAE= 5.12 years. Recently, majority methods estimated age by regression, such as linear regression [17], SVR [10], PLS [20], and CCA [3]. Due to the ability of handling multiple tasks, PLS and CCA hold the best performance in the literature [3].

Among existing methods, BIF+CCA [3] was almost the best method for practical applications in terms of accuracy and speed. Generally speaking, we can see BIF+CCA as a 3-layered network composed by convolutional, pooling and full connected layers. Deep neural network has also been used for age estimation [9] and gender classification [21], but its potentials were not worked out completely. As described in the previous section, the biggest contribution of this paper is combing some strategies of traditional methods into CNN to improve the state-of-the-art.

This work is mainly inspired by CNN and the three tricks in traditional methods: multi-scale analysis, local aligned face patch, and facial symmetry. Compared with the proposed network, a similar multi-scale CNN was proposed for scene labeling [22] not long ago, but the multi-scale analysis in [22] was mainly used in testing stage while ours is used both in training and testing stages. Local aligned patch has succussed in many methods [23] for unconstrained face recognition problem. And facial symmetry is also a widely used trick to deal with pose problem [24] or reduce the dimension of face image and augment the database [25].

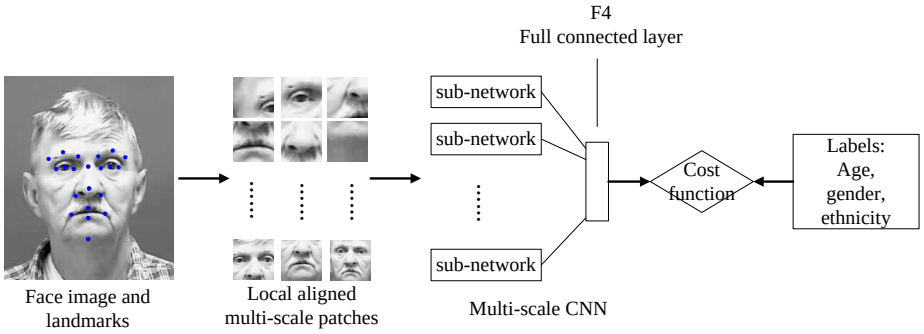

Fig. 1. The structure of the proposed network. The input face image is cropped into many local aligned patches. All patches are fed to the multi-scale convolutional network. The response of each patch are combined at the full connected layer to estimate the age, gender and ethnicity.

3 Multi-scale Convolutional Network

The structure of the proposed network is shown in Fig. 1, which includes many sub-networks for each patch. The details of the network are described in the following contents.

3.1 Local Aligned Face Patches

Facial landmarks are important for good face recognition algorithms, especially for unconstrained face recognition problem. Based on the precise facial landmarks, one can correct the pose of face image or build pose robust face descriptors. The most simple and effective method to use landmarks is local alignment. By cropping face patch around each landmark, we can get some patches aligned in the local coordinate system. For different face images, these patches have the same semantics, which are better than holistic face to learn high level tasks.

Due to the success of landmarks in face recognition, we crop face image into many local aligned patches as the input of our network. Given a face image, we first localize 21 facial landmarks by ASM [26]. The positions of the detected landmarks are shown in Fig. 2. According to facial symmetry, we group the landmarks into 13 pairs. The index of the landmark pairs are: (1, 2), (3, 4), (5, 6), (7, 8), (9, 10), (11, 12), (13, 14), (15, 16), (17, 17), (18, 18), (19, 19), (20, 20), (21, 21). For those points (17-20) on the middle line of face image, they form pairs to themselves.

In accordance with other papers, all color images are first converted into gray, because the color information is unstable and useless for age estimation. Before cropping image patches, the distance of reference landmarks 17 and 19 (*i.e.,* scale) of all face images are normalized to 60, 42, 30, 22 pixels in 4 scales. All landmarks are transformed along with the normalization of images. On the

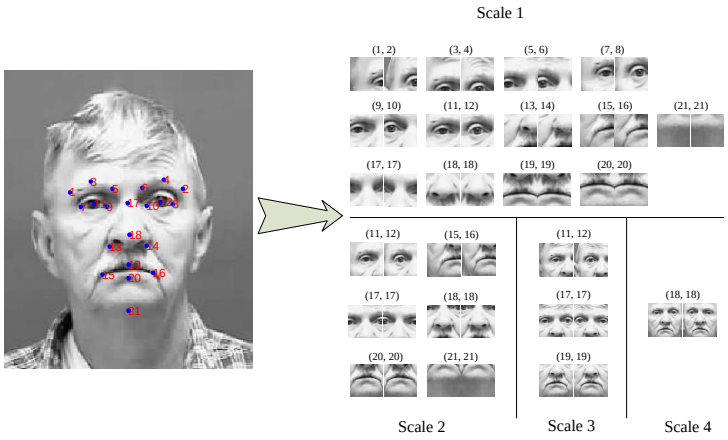

Fig. 2. 23×2 multi-scale patches cropped from a face image based on its corresponding landmarks. The resolution of all patches are 48×48. The patches from the right half of face are mirrored to augment the database.

normalized images, several 48×48 patches are cropped in 4 scales (differing by half-octave) by taking the landmarks as center. The number of landmark pairs used in 4 scales are 13, 6, 3 and 1 from small to large scale, thus giving 23 multi-scale patch pairs. For the landmarks in the right half of face, the patches are mirrored to be consistent with the the left half. In this way, we can get $23 \times 2 = 46$ local aligned, multi-scale image patches for a face image, which are shown in Fig. 2. The patches capture the appearance of the face image in multi-scale and are robust to rigid and non-rigid deformations.

3.2 Convolutional Network for Age Estimation

Fig. 1 shows the architecture of the proposed network, the details of each layer is described in Fig. 3. For the 23 groups of image patches, we create 23 sub-networks to process them respectively and fuse their responses in the final full connected layer to estimate the age. This structure has two benefits: 1) 23 sub-networks can learn the particular features for each patch; 2) The final layer connects all sub-networks together, which can make them mutually complementary. Note that the parameters of the 23 sub-networks and the final layer in Fig. 1 and Fig. 3 are optimized in a whole process.

The sub-network for each patch is composed by a convolutional layer, a max pooling layer and a local layer (Locally-connected layer with unshared weights). The number of channels of the convolutional, pooling and local layers are both 16. Before convolution the input are padded by zero values, therefore the output have the same size with input. The filter size of C1 layer is 7×7 and the filter size of L3 layer is 3×3. ReLU neuron [27] is used as activation function for C1 and L3. The stride of S2 is 3 that means it down-samples the feature maps

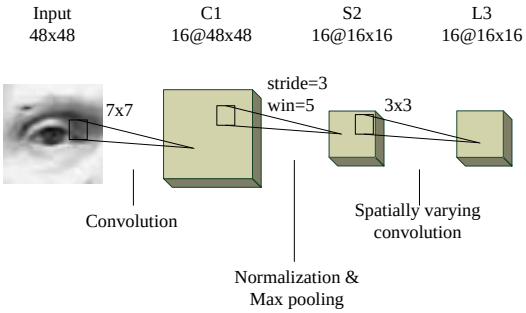

Fig. 3. The structure of sub-network for each patch. The input of sub-network are face patches and the output are sent to the F4 layer in Fig. 1.

from 48×48 to 16×16 and S2 includes a cross-channel normalization unit. The output of the sub-network (L3) has $16 \times 16 \times 16 = 4096$ dimensions. Therefore the input of F4 has $4096 \times 23 = 94208$ dimensions. F4 uses square difference as cost function, so it can be seen as a linear regression layer. In practice, we should pay attention to the magnitude of F4's output and the target (age). Generally, we need introduce a scale factor to make them in the same order of magnitude. The network are optimized by Stochastic Gradient Descent (SGD).

Besides the learnable parameters, the proposed network has an extra local layer when compares with BIF+CCA [3]. As we know, convolution is appropriate to capture the statistics over the whole image, but the statistical properties of face image are not stationary with respect to the location in image. Thus we add a local layer to model the spatially varying statistics in the high level of network. The following experiments will illustrate the improvement produced by this layer.

3.3 Multi-task Learning

Age, gender and ethnicity are three close related traits of human. When we estimate them from face images, the three traits interact each other. [28] first estimated the gender and ethnicity of a face image and then sent it to a gender-ethnicity group based age estimator, which obtained good results. [20] and [3] used PLS and CCA to estimate the three traits simultaneously and obtained better results than previous methods. Jointly estimating age, gender and ethnicity has two advantages: 1) By sharing the model between the three tasks can improve the speed of learning and inference; 2) Multi-label can supply more information of the database to regularize the network during training.

Based on the proposed network, we extend it by a multi-task loss function to estimate age, gender and ethnicity of face image jointly. The output of F4 layer in Fig. 1 is modified from 1 to 3 dimensions, in which each dimension is corresponded to age, gender and ethnicity respectively. Our multi-task loss function is composed by three terms: a square loss for age, binomial deviance

loss [29] (also known as cross-entropy) for gender and ethnicity. For ethnicity classification, we only report accuracy for the Black and White because the majority images of MORPH Album 2 are White and Black (96%). The loss function is

$$\begin{aligned} J =&(C(X,W)_{age} - L_{age})^2 + \alpha \ln(e^{-2C(X,W)_{gender}L_{gender}} + 1) + \\ &\beta \ln(e^{-2C(X,W)_{ethnicity}L_{ethnicity}} + 1), \end{aligned} \tag{1}$$

where $C(X, W)$ denotes the function of the network. X is the input face image. W is the parameters of the network. The subscripts "age", "gender" and "ethnicity" denote the 3 dimensions of output. L is the 3 dimensional label of training set. $L_{gender} \in (-1, 1)$, -1 denotes Male and 1 denotes Female. $L_{ethnicity} \in (-1, 1)$, -1 denotes Black and 1 denotes White. α and β are hyper-parameters to tune the importance of each term.

Because Eqn.(1) is derivable, the objective can be easily optimized by SGD too. If need deal with multiple ethnicity classification on other databases, we can use softmax regression and negative log-likelihood as loss function.

4 Experiments

Large data is needed to train a good neural network, therefore we conduct the experiments on MORPH Album 2, which is the only large aging database we know. First, the information and setup of MORPH is described. Then, four networks with different architecture are compared to illustrate the superiority of the proposed network. And the benefits of multi-scale analysis and local alignment are verified by the results. Finally, the proposed network is compared with state-of-the-art methods and an efficient multi-task network is presented.

4.1 Database and Setup

MORPH Album 2 contains about 55,000 face images of more than 13,000 subjects. The capture time spans from 2003 to 2007. Age ranges from 16 to 77 years. Although it is a good and large database, the distributions of gender and ethnicity are uneven. The Male-Female ratio is about 5.5 : 1 and the White-Black ratio is about 4 : 1. Except for White and Black, the proportion of other ethnicity is very low (4%).

To use the database effectively, we follow the previous way [28] to preprocessing the database and split it into three non-overlapped subsets S_1, S_2 and S_3 randomly. First, all images in MORPH are processed by a face detector [30]. Because MORPH contains some non-face images (*e.g.,* tattoo), they are removed from the database after this step. The number of face images in the processed database is 55244. Then the facial landmarks of face images are localized by ASM [26] and local aligned patches are cropped based on the landmarks described in Section 3.1. We construct the $S1$, $S2$ and $S3$ subsets by two rules: 1) Making Male-Female ratio ≈ 3; 2) Making White-Black ratio $= 1$. The

Table 1. The information of the pre-processed MORPH Album 2 and S1, S2, S3 subsets.

	Black			White			Other
Male	S1: 4012	S2: 4012	S3: 28835	S1: 4012	S2: 4012	S3: 0	S3: 1845
Female	S1: 1305	S2: 1305	S3: 3166	S1: 1305	S2: 1305	S3: 0	S3: 130

information of the subsets are shown in Table 1. In all experiments, the training and testing are repeated in two times: 1) training on $S1$, testing on $S2 + S3$ and 2) training on $S2$, testing on $S1 + S3$. The performance of the two experiments and their average are reported. For age and gender estimation, all images in Table 1 are used. For ethnicity classification, the images in "Other" column are neglected.[1]

4.2 Age Estimation

In the age estimation experiments, we will illustrate the advantages of the proposed network in three aspects: architecture, multi-scale analysis, and local alignment.

As described in Section 3.1, $23 \times 2 = 46$ multi-scale patches are generated for every face image. We call those not mirrored patches as left-patches (marked by blue box in Fig. 4), and the mirrored patches as right-patches (marked by red box). In the training stage, the left-patches and right-patches can be seen as a way to double the training set. In the test stage, we can get two predicts based on the left-patches and right-patches respectively and fuse the two predicts by average. The test process is shown in Fig. 4.

Architecture The architecture of neural network determines the capacity of the model. How to choose a good architecture is a problem specific task and is also affected by the scale of training data. In this section, we compare 4 networks with various architectures from shallow to deep. The architectures for comparison are as follows.

1. C-P-F: convolution + max pooling + full connection;
2. C-P-L-F (proposed): convolution + max pooling + local layer (Locally-connected layer with unshared weights) + full connection;
3. C-P-C-F: convolution + max pooling + convolution + full connection;
4. C-P-C-P-L-F: convolution + max pooling + convolution + max pooling + local layer + full connection.

"C-P-L-F" is the proposed architecture which has been described in Section 3.2. In all networks, the number of filters are both 16. The number of training epoch is set to 30. Before training, all images subtract the mean value over the training set from each pixel.

[1] The detailed evaluation protocols and facial landmarks can be downloaded from http://www.cbsr.ia.ac.cn/users/dyi/agr.html.

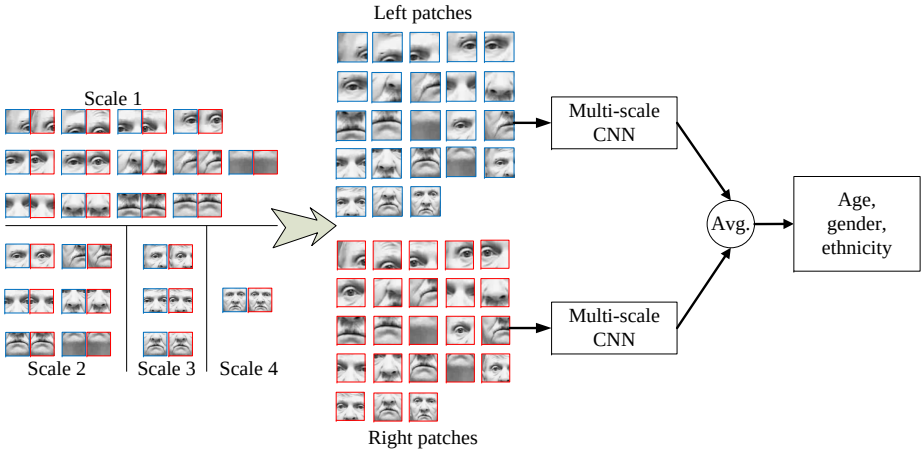

Fig. 4. The test process of the proposed network. For a face image, $23 \times 2 = 46$ multi-scale patches are divided into two groups "left patches"(blue box) and "right patches"(red box) and their predicts are fused by average.

Table 2. Comparison of 4 networks with different architectures.

Architecture	Training Set	Test Set	Age MAE	Average MAE
C-P-F	S1	S2 + S3	3.75	3.71
	S2	S1 + S3	3.66	
C-P-L-F (proposed)	S1	S2 + S3	3.63	**3.63**
	S2	S1 + S3	3.63	
C-P-C-F	S1	S2 + S3	3.94	3.93
	S2	S1 + S3	3.91	
C-P-C-P-L-F	S1	S2 + S3	3.85	3.78
	S2	S1 + S3	3.71	

The performance of age estimation of above 4 networks are shown in Table 2. From the table we can see many interesting results. The proposed architecture "C-P-L-F" has lower error than other compared architectures. Using a local layer before the full connected layer can always improve the performance, *i.e.,* "C-P-L-F" is better than "C-P-F" and "C-P-C-P-L-F" is better than "C-P-C-F", which illustrate the importance of the local layer. Deeper architectures ("C-P-C-F" and "C-P-C-P-L-F") have worse performance than the shallow ones ("C-P-F" and "C-P-L-F"). The reason may be that the potentials of deeper architecture are not developed by the current scale of training data. In the following experiments, we will set the architecture as "C-P-L-F".

Multi-scale analysis As shown in Fig. 2, the patches are generated in 4 scales. And the number of patches in each scale are: 13, 6, 3, 1. To illustrate the advantage of multi-scale analysis in our network, we evaluate the performance of

Table 3. The performance of networks in single scale and multi-scale.

Scale	#Patches	Training Set	Test Set	Age MAE	Average MAE
Scale 1	13	S1	S2 + S3	3.84	3.87
		S2	S1 + S3	3.89	
Scale 2	6	S1	S2 + S3	4.39	4.41
		S2	S1 + S3	4.43	
Scale 3	3	S1	S2 + S3	4.44	4.36
		S2	S1 + S3	4.27	
Scale 4	1	S1	S2 + S3	5.53	5.45
		S2	S1 + S3	5.37	
Multi-scale (proposed)	23	S1	S2 + S3	3.63	**3.63**
		S2	S1 + S3	3.63	

Table 4. The performance of two networks trained on aligned and non-aligned patches.

	Training Set	Test set	Age MAE	Average MAE
Non-aligned	S1	S2 + S3	3.87	3.79
	S2	S1 + S3	3.70	
Aligned (proposed)	S1	S2 + S3	3.63	**3.63**
	S2	S1 + S3	3.63	

each scale respectively. Except for the number of sub-networks, the architecture for each scale is as same as the multi-scale version. The results of each scale and multi-scale are shown in Table 3. Small scale generally has better performance than large scale. There are three possible reasons for this phenomenon: first, patches in small scale contain more texture information which are close related to human age; second, small scale patches are better aligned than large scale patches; third, the number of patches in small scale is more than large scale.

Although different scales have different performance, they are very complementary to each other. When fusing the 4 scales together, the MAE is reduced significantly from 3.87 to 3.63.

Local Aligned Patches Besides of the multi-scale analysis, local alignment is our another contribution. Here, we conduct an experiment to verify the power of local alignment. For fairness, we crop patches for all face images again but based on a mean shape, which is generated by averaging the landmarks of all face images in MORPH Album 2. Because the mean shape cannot be accurate for every face image, the cropped patches are not aligned well. Then we use these not well aligned patches to train a network for comparison. We call these two networks as aligned network and non-aligned network. The two networks have the same structure and their inputs are both $23 \times 48 \times 48$ dimensions. Some aligned and non-aligned patches of a face image are shown in Fig. 5.

Table 4 shows the performance of the aligned and non-aligned networks. Apparently, the MAE of the aligned network is lower than the non-aligned one. Due to the non-aligned patches have more variations in translation, the non-

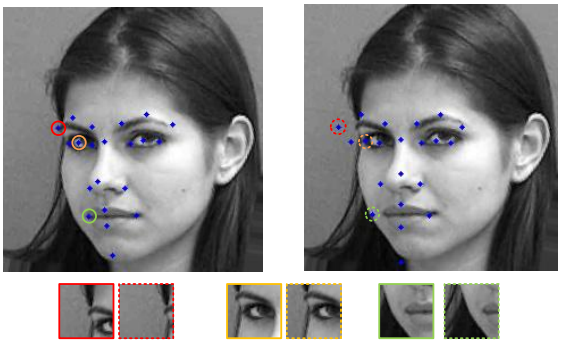

Fig. 5. Left: A face image and its corresponding landmarks. Right: The face image and the positions of the mean shape. Bottom: Some aligned and non-aligned patches of the face image. For abbreviation, just three patch pairs are shown.

aligned network should pay more attention to learn the translation invariant feature, which reduces the performance of age estimation. On the contrary, the input patches of the aligned network are already aligned well, therefore the network can focus on age estimation better. Moreover, the non-aligned network need 35 epoches to converge, which is slightly more than 30 epoches of the aligned network.

The improvements of each module in the proposed method has been verified in the above paragraphs, including architecture, multi-scale analysis and local alignment. The overall performance of the proposed method is compared with state-of-the-art methods in Table 5. The comparison is done on MORPH Album 2 too, the largest public database for age estimation. From the last column of Table 5, we can see the existing lowest MAE= 3.98 is obtained by BIF+KCCA [3]. The proposed multi-scale CNN reduces the MAE to 3.63 significantly, and the relative reduction is 8.8%. From the results, we can find another interesting thing for BIF, that is the kernel methods are consistently better than their linear version, such as KCCA > (r)CCA, KPLS > PLS and KSVM > LSVM. The observation indicates that the "convolution + pooling + kernel correlation" architecture is better than "convolution + pooling + linear transformation" in age estimation. In our intuition, the architecture of the proposed network, "C-P-L-F"(see Section 4.2), is very similar to "BIF+kernel method". By comparing Table 5 and Table 3, we can see that even using a single scale our MAE= 3.87 is still lower than BIF+KCCA.

4.3 Joint Estimation of Age, Gender and Ethnicity

Finally, we train a multi-task network to estimate age, gender and ethnicity simultaneously from face image using the method described in Section 3.3. The training process is similar to the previous experiment, architecture="C-P-L-F", epoch= 30, and optimized by SGD, except for the label of training samples including extra gender and ethnicity information. The size of α and β in Eqn.(1)

Table 5. The results of estimating age, gender and ethnicity on MORPH Album 2 and comparison of the proposed method to state-of-the-art methods.

Method	Tr. Set	Test set	Gender Accu.	Ethnicity Accu.	Age MAE	Avg. MAE
CNN (Single-Task)	S1	S2 + S3	-	-	3.63	**3.63**
	S2	S1 + S3	-	-	3.63	
CNN (Multi-Task)	S1	S2 + S3	98.0%	99.1%	3.72	**3.63**
	S2	S1 + S3	97.8%	98.1%	3.54	
Baseline CNN [9][2]	S1	S2 + S3	-	-	4.64	4.60
	S2	S1 + S3	-	-	4.55	
BIF+CCA [3]	S1	S2 + S3	95.2%	97.8%	5.39	5.37
	S2	S1 + S3	95.2%	97.8%	5.35	
BIF+rCCA [3]	S1	S2 + S3	97.6%	98.7%	4.43	4.42
	S2	S1 + S3	97.6%	98.6%	4.40	
BIF+KCCA [3]	S1	S2 + S3	98.5%	98.9%	4.00	3.98
	S2	S1 + S3	98.4%	99.0%	3.95	
BIF+PLS [20]	S1	S2 + S3	97.4%	98.7%	4.58	4.56
	S2	S1 + S3	97.3%	98.6%	4.54	
BIF+KPLS [20]	S1	S2 + S3	98.4%	99.0%	4.07	4.04
	S2	S1 + S3	98.3%	99.0%	4.01	
BIF+3Step [28]	S1	S2 + S3	98.1%	98.9%	4.44	4.45
	S2	S1 + S3	97.9%	98.8%	4.46	
BIF+LSVM [3]	S1	S2 + S3	-	-	5.06	5.09
	S2	S1 + S3	-	-	5.12	
BIF+KSVM [3]	S1	S2 + S3	-	-	4.89	4.91
	S2	S1 + S3	-	-	4.92	
CPNN [31]	10-fold CV		-	-	-	4.87

just has a little influence to the performance of the multi-task network, so we set them as 1 in the experiment. The performance of the multi-task network is list in the second row of Table 5. When jointly optimizing with gender and ethnicity classification, the average MAE of age estimation is as same as the single-task network, which are both 3.63 years. And meanwhile, the accuracies of gender and ethnicity classification are comparable to state-of-the-art. Being consistent with other works, the accuracy of ethnicity classification is higher than gender. This may be a clue to design novel multi-task networks in the future.

Compared with the proposed method, the most competitive methods are BIF+KCCA and BIF+KPLS, but their test speed are slow. On a Intel Core 2 CPU@2.1GHz, excluding the time of feature extraction, the test time of KCCA and KPLS are 72515.6 and 72516.4 seconds [3]. On a Tesla K20 GPU, the test time of our network on the whole test set is 87 seconds (44610 samples, 2ms/sample), which is faster than kernel methods significantly. On a Intel Core i3-2370M@2.4GHz (single thread), the test time of our CPU version is 8900 seconds (200ms/sample), which is 100× slower than the GPU version, but still

[2] This 5-layered baseline CNN is implemented by us according to the description in [9]. The dimension of face images to train and test the network is 64×64.

faster than KCCA and KPLS significantly. Note that our test time is measured from inputting the aligned patches to outputting the results.

5 Conclusions

We proposed a novel age estimation method based on CNN in this paper. Compared with the state-of-the-art BIF based methods, our method achieved significant lower error on MORPH Album 2 database due to its deeper structure and learnable parameters. To apply CNN effectively in age estimation, we carefully designed the architecture of the network and combined two important tricks into the network. Extensive experiments illustrated the improvements brought by our design principles, including "C-P-L-F" architecture, multi-scale analysis and local aligned patches. Furthermore, we constructed a novel loss function for age, gender and ethnicity joint estimation and trained a multi-task network. Experiments showed that our multi-task network achieved the same MAE with the single-task network and achieved high accuracies in gender and ethnicity classification at the same time. Future work will focus on how to design more reasonable multi-task architectures for age, gender and ethnicity joint estimation.

Acknowledgment

This work was supported by the Chinese National Natural Science Foundation Projects #61105023, #61103156, #61105037, #61203267, #61375037, #61473291, National Science and Technology Support Program Project #2013BAK02B01, Chinese Academy of Sciences Project No. KGZD-EW-102-2, and AuthenMetric R&D Funds. The GPU was donated by NVIDIA.

References

1. Kwon, Y.H., da Vitoria Lobo, N.: Age classification from facial images. In: Computer Vision and Pattern Recognition, 1994. Proceedings CVPR '94., 1994 IEEE Computer Society Conference on. (1994) 762–767
2. Lanitis, A., Draganova, C., Christodoulou, C.: Comparing different classifiers for automatic age estimation. Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on **34** (2004) 621–628
3. Guo, G., Mu, G.: Joint estimation of age, gender and ethnicity: Cca vs. pls. In: FG. (2013) 1–6
4. Rawls, A., Ricanek, K.: Morph: Development and optimization of a longitudinal age progression database. In: COST 2101/2102 Conference. (2009) 17–24
5. Basak, D., Pal, S., Patranabis, D.C.: Support vector regression. Neural Information Processing-Letters and Reviews **11** (2007) 203–224
6. Geladi, P., Kowalski, B.R.: Partial least-squares regression: a tutorial. Analytica chimica acta **185** (1986) 1–17
7. Hardoon, D.R., Szedmak, S., Shawe-Taylor, J.: Canonical correlation analysis: An overview with application to learning methods. Neural computation **16** (2004) 2639–2664

8. Daugman, J.G.: Complete discret 2d gabor transforms by neural networks for image analysis and compression. IEEE Trans. ASSP **36** (1988) 1169–1179
9. Yang, M., Zhu, S., Lv, F., Yu, K.: Correspondence driven adaptation for human profile recognition. In: CVPR. (2011) 505–512
10. Guo, G., Mu, G., Fu, Y., Huang, T.S.: Human age estimation using bio-inspired features. In: CVPR. (2009) 112–119
11. Kwon, Y.H., da Vitoria Lobo, N.: Age classification from facial images. Computer Vision and Image Understanding **74** (1999) 1–21
12. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. In: ECCV98. Volume 2. (1998) 484–498
13. Geng, X., Zhou, Z.H., Smith-Miles, K.: Automatic age estimation based on facial aging patterns. IEEE Trans. Pattern Anal. Mach. Intell. **29** (2007) 2234–2240
14. Gao, F., Ai, H.: Face age classification on consumer images with gabor feature and fuzzy lda method. In: Advances in biometrics. Springer (2009) 132–141
15. Gunay, A., Nabiyev, V.V.: Automatic age classification with lbp. In: Computer and Information Sciences, 2008. ISCIS'08. 23rd International Symposium on, IEEE (2008) 1–4
16. Yan, S., Liu, M., Huang, T.S.: Extracting age information from local spatially flexible patches. In: Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on, IEEE (2008) 737–740
17. Fu, Y., Huang, T.S.: Human age estimation with regression on discriminative aging manifold. Multimedia, IEEE Transactions on **10** (2008) 578–584
18. Cao, D., Lei, Z., Zhang, Z., Feng, J., Li, S.Z.: Human age estimation using ranking svm. In: CCBR. (2012) 324–331
19. Herbrich, R., Graepel, T., Obermayer, K. In: Large margin rank boundaries for ordinal regression. MIT Press, Cambridge, MA (2000)
20. Guo, G., Mu, G.: Simultaneous dimensionality reduction and human age estimation via kernel partial least squares regression. In: CVPR. (2011) 657–664
21. Duffner, S.: Face image analysis with convolutional neural networks. PhD thesis (2008)
22. Farabet, C., Couprie, C., Najman, L., LeCun, Y.: Learning hierarchical features for scene labeling. IEEE Trans. Pattern Anal. Mach. Intell. **35** (2013) 1915–1929
23. Chen, D., Cao, X., Wen, F., Sun, J.: Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification. In: CVPR. (2013) 3025–3032
24. ul Hussain, S., Wheeler, Napolon, T., Jurie, F.: "Face recognition using local quantized patterns". In: Proc. British Machine Vision Conference. Volume 1. (2012) 52–61
25. Li, S.Z., Yi, D., Lei, Z., Liao, S.: The casia nir-vis 2.0 face database. In: CVPR Workshops. (2013) 348–353
26. Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J.: "Active shape models: Their training and application". CVGIP: Image Understanding **61** (1995) 38–59
27. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS. (2012) 1106–1114
28. Guo, G., Mu, G.: Human age estimation: What is the influence across race and gender? In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on. (2010) 71–78
29. Friedman, J., Tibshirani, R., Hastie, T.: The Elements of Statistical Learning : Data Mining, Inference, and Prediction. Springer Series in Statistics. Springer-Verlag, New York (2009)

30. Viola, P., Jones, M.: "Rapid object detection using a boosted cascade of simple features". In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Kauai, Hawaii (2001)
31. Geng, X., Yin, C., Zhou, Z.H.: Facial age estimation by learning from label distributions. IEEE Trans. Pattern Anal. Mach. Intell. **35** (2013) 2401–2412