Unleash the Black Magic in Age: a Multi-task Deep Neural Network Approach for Cross-age Face Verification

Xiaolong Wang[1], Yin Zhou[1], Deguang Kong[1], Jon Currey[1], Dawei Li[1], Jiayu Zhou[2]
[1] Samsung Research America, Mountain View, CA 94043, USA
[2] Michigan State University, East Lansing, MI 48824, USA

***Abstract*— Facial aging is a complicated process which usually affects the facial appearance (*e.g.*, wrinkles). Variations of facial appearance pose a big challenge to the automatic face recognition problem. How to eliminate the influence of aging factors to the verification performance is a very challenging problem. Multi-task learning has provided a principled framework for jointly learning multiple related tasks to improve generalization performance. In this paper, we leverage this powerful technique to improve the task of cross-age face verification. We present an end-to-end learning framework for cross-age face verification by designing a multi-task deep neural network architecture that exploits the intrinsic low-dimensional representation shared between the tasks of face verification and age estimation. We show that the algorithm effectively balances feature sharing and feature exclusion between the two given tasks. We evaluate the proposed framework on two standard benchmarks. Experimental results demonstrate that our algorithm has significant improvement over the state-of-the-art (2.2% EER on MORPH and 7.8% EER on FG-NET, by more than 50.0% and 59.70% performance gain respectively).**

I. Introduction

As an important and challenging problem, face verification has been widely investigated in computer vision and biometric areas for many years. Face recognition and retrieval with age gaps have a wide range of applications. For example, it can be used to help find missing people, especially to help identify these trafficking children after a long period of time. This is especially useful in forensic areas [5], [13]. It also has a wide range of other applications including security, law-enforcement, visual surveillance and human computer interaction, etc. In these applications, age information is usually the major influential factor for face verification.

Although significant progress has been made in face recognition [37], [39], [40], [52], cross-age face verification still remains a challenging area and gains increasing attentions. The intrinsic challenge lies at the conflicting patterns between age and identity.

Images taken with age variations usually lead to huge difference in facial appearance. Intuitively, to perform cross-age face recognition task, one direct way is to synthesize the input facial image to the target age by facial modeling [33]. This generative solution usually faces a dilemma: It is very difficult to genuinely synthesize and simulate the unpredictable aging progress accurately since the process of the aging patterns is different between different individuals and are affected by many factors (e.g., living environment, living environment, health condition, lifestyle). To address these challenges, age-invariant facial features are widely adopted.

The key idea of this kind of approaches is to decompose aging/identity patterns separately and extract age-invariant identity descriptors [7], [13]. Intuitively, the learning framework is, in some sense, equivalent to a multi-task learning framework with two joint tasks: face verification and age estimation. The only difference is that, instead of exploring two tasks' similarities, the framework learns the conflicting relationship and encourages feature decomposition by the nature. However, their learning framework requires consolidated input/output formats for all tasks. Unfortunately, face verification task takes image pairs $\{0, 1\}$ as input/output while age estimation takes single image/integer as input/output. We can employ an ad hoc scheme to replace age estimation by age verification task, which shares the same input/output format as face verification. But age estimation is always the most natural choice since the age verification result cannot accurately characterize facial aging cues in details. This intuition suggests that using age estimation task directly can further improve verification. The intuition is later validated through our experimental results. Meanwhile, recent advances in convolutional neural nets dramatically improved the state-of-the-art in many facial analysis domains. Now the similar question again looms on cross-age identity verification: how can we leverage CNN for this virgin domain?

Following the above observation, our proposed framework is motivated by the insight: In order to further improve cross-age verification, how can we design a multi-task deep learning framework to learn aging-invariant feature, and graft face identity/age estimation tasks together into one integrated framework? To this end, we design a new jointed deep neural network to learn facial features. In the whole framework, identity verification is designed as the main task where age estimation is set as the auxiliary task. In this framework we construct a Siamese network [6] where two coupled deep convolutional neural networks share the same parameters. For identity training, the contrastive loss is used for verification task and cross-entropy loss is used for age estimation task.

In this work, we explore a deep learning architecture along with cross-age verification and demonstrate its power for dealing with this newly investigated domain. Further, our experiments show an impressive result that, although we design the network only seeking for age-invariant identity features, the proposed deep neural network can well

978-1-5090-4023-0/17 $31.00 © 2017 IEEE
DOI 10.1109/FG.2017.75

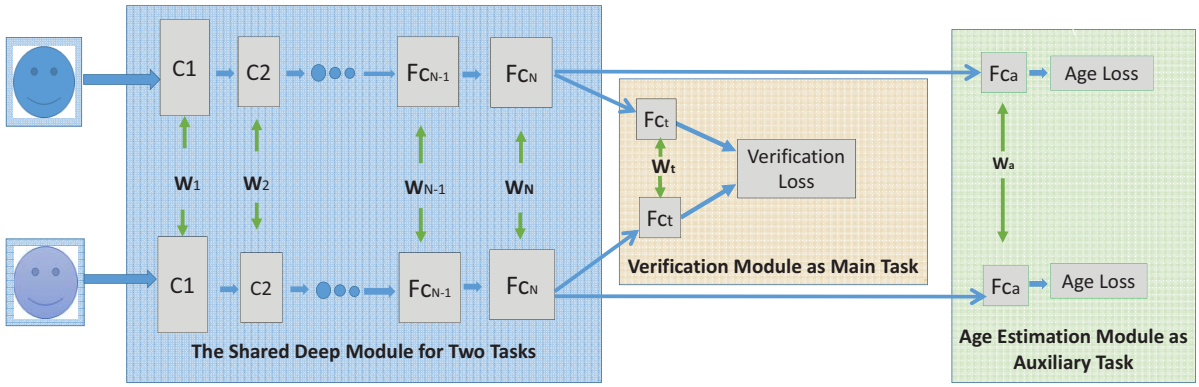


Fig. 1. The proposed joint deep network architecture is based on the Siamese deep neural network for both face recognition and age estimation. It consists of two pipelines/tasks: (a) face recognition pipeline (b) age estimation pipeline. Given the pairwise images, the face recognition pipeline processes the features using deep neural network (DNN) encoded as $W^F = [W_1, W_2, \cdots, W_n]$ and feed them into the last layer encoded as W_t before minimizing the face recognition error given two pairwise images using a constractive loss (Eq.3). For each image, the age estimation pipeline processes the features using deep neural network encoded as $W^A = [W_1, W_2, \cdots, W_n]$ and feed them into the last layer encoded as W_a before minimizing the age estimation error for each image using a cross-entropy loss (Eq.4). In the figure, we assume the DNN structure is shared among the two pipelines/tasks, i.e., $W^F = W^A$. In our framework, the layers $Fc_1 - Fc_N$ are directly adapted from VGG. Fc_N indicates Fc_7. Besides that, we designed the new layers Fc_t and Fc_a.

decompose identity/age and thus learn high genuine aging estimation features simultaneously. Namely, this network is also capable to serve as a multi-task framework for both age estimation and identity verification. We have applied the proposed framework on two widely used face-aging benchmark datasets: MORPH [35] and FG-NET [1]. The obtained verification performance outperforms previous algorithms by a large margin. Besides that, our scheme also achieves comparable performance in the auxiliary task-age estimation. The proposed method is essentially different compared to any other previous cross-age face verification and age estimation frameworks. Experimental results together with the experimental analysis clearly indicate the benefit of building two sub-tasks for age-related face image analysis.

II. Related Work

A. Cross-age face recognition

In recent years, researchers become much more interested in cross-age face recognition topics [6], [12], [23], [33], [41]–[43]. In general, we can divide these works into two categories. The first category can be classified into generative models which is to transform face image for characterizing aging process. Ramanathan and Chellappa [33] proposed a craniofacial growth model to simulate the facial appearance changes across years. They translated the age-based anthropometric constraints on facial proportions onto the facial growth parameters. They had tested the proposed approach on the database where people's ages are under 18 years old. Park et al. [28] proposed a 3D aging modeling technique to compensate the age variations. Their model also considered the variations caused by view changes [42]. Suo et al. [42] devised a compositional and dynamical model to represent faces in each age group. This model is based on a hierarchical And-Or graph where different nodes represent different facial details. From these works, we observe that the major difficulty for modeling aging process is caused by the complexity of human aging process where it is very difficult to capture aging information accurately.

The second general category can be categorized into discriminative methods. The general idea is to devise robust feature or optimize classifier to improve the verification performance. One representative work is proposed by Ling et al. [24]. In this work, they used gradient orientation pyramids (GOP) as feature representation and support vector machines (SVMs) as the classifier for verification. Li et al. [23] described face images by fusing (scale invariant feature transform (SIFT) and multi-scale local binary patterns (MLBP) features together from densely sampled facial regions. Multi-feature discriminant analysis (MFDA) is used to reduce the feature dimension. To improve the verification performance, random sampling methods were also applied to the feature space. Multiple LDA-based classifiers are fused together to obtain the final verification result. Similarly, Sungatullina1 et al. [41] applied discriminative learning approach on the fusion feature space (SIFT, LBP and GOP) to project different types of local features into a latent discriminative subspace to minimize the samples' variation of the same subject. Subspace feature learning with feature fusion is also applied in recent works for cross-age face recognition [12], [13]. A coding framework based on measuring relative similarities to the reference data is also proposed to solve this kind of problem [5]. Du and Liang [7] developed an alternating greedy coordinate descent (AGCD) algorithm to solve the face and age verification together. Their results illustrate that utilizing aging information is helpful in improving the verification performance.

B. Age estimation

Facial image age estimation has been studied for a long time [3], [4], [9], [11], [15], [16], [20], [22], [34], [36], [46]–[48]. Age estimation can be categorized into aging feature representation and predication sub-problems. For the aging feature representation, previous works usually want to use the shape model to simulate the aging process [20]. However, the limitation of this kind of work is that predicting the age based on the shape information is not accurate. Compared to shape information, representing aging information using texture information attracts more interests [15]. Among them, one of the efficient aging feature representations is biologically inspired features (BIF), which demonstrates significant improvement compared to traditional features (LBP, SIFT, Gabor, HOG and sparse features [**?**], [2]) and has been widely used in this area [10], [17], [27]. Recently, with the popularity of deep neural network, the performances have achieved significant improvement over the past few years [21], [45], [51]. The results obtained in these frameworks have also motivated us to build a unified network for verification and age estimation.

III. Cross-age Face recognition via Multi-task DNN

A. Overview of the approach

Given the diversified images from different persons at different ages, our goal is to automatically identify whether the faces are from the same person and infer the age of the person as well. This is a challenging task that requires good performances on each individual one. The key idea of our approach is to leverage deep nerual network to obtain better image feature representations for both pairwise images and individual image, and further take advantage of multi-task learning approach to perform the two tasks simultaneously. As is illustrated[1] in Fig. 1, given a pair of images (X_i, X_j), the proposed framework consists of two pipelines:

• Face recognition pipeline: it learns a deep neural network encoded as model parameters $W^F = [W_1^F, W_2^F, \cdots, W_n^F]$ before feeding the new feature representation (Z_i, Z_j) into the last layer quantified by parameter W_t and get prediction results (y_i, y_j). In the final layer, the constractive loss (Eq.3) is applied to minimize the face recognition error given two pairwise images. The trained model parameter is given by: $\theta_F = [W^F, W_t]$, *i.e.,*

$$\mathrm{T}ask1: \left.\begin{array}{c} X_i \overset{W^F}{\rightarrow} Z_i \overset{W^t}{\rightarrow} y_i \\ X_j \overset{W^F}{\rightarrow} Z_j \overset{W^t}{\rightarrow} y_j \end{array}\right\} \text{Constrative Loss}$$

• Age recognition pipeline: it learns a deep neural network encoded as $W^A = [W_1^A, W_2^A, \cdots, W_n^A]$ for each image X_i before feeding the new feature representation $\tilde{Z}_i$ into the last layer quantified by parameter W_a and get prediction result $\tilde{y}_i$. In the final layer, the cross-entropy loss (Eq.4) is applied to minimize the age recognition error given each image. The trained model parameter is given by: $\theta_A = [W^A, W_a]$, *i.e.,*

$$\mathrm{T}ask2: \quad X_i \overset{W^A}{\rightarrow} \tilde{Z}_i \overset{W^a}{\rightarrow} \tilde{y}_i; \quad \text{Cross-entropy Loss.}$$

The two pipelines are viewed as two tasks where face recognition is viewed as the main task and are performed jointly. Further the network structure W^F and W^A can be shared between the two tasks, *i.e.,* $W^F \approx W^A$. In the testing phrase, we only need to feed the new testing image pairs into the learned network $\{\theta_F, \theta_A\}$ and get both the face detection and age identification results.

B. Why Multi-task learning?

The idea of multi-task learning [14], [55]–[57] is to simultaneously learn a set of related learning tasks due to the fact that the feature learned from a certain task may also be useful for other task [54] as is evident in tracking [53], facial landmark detection [54]. The joint learning process would enforce the learning of one task to bias and be biased by other tasks, and thus allow useful predictive knowledge to be transferred among related tasks. Specifically, suppose that we have a collection of T supervised learning tasks. We are given the training data of T tasks $\mathscr{D} = \{\mathscr{D}_1, \ldots, \mathscr{D}_T\}$, where $\mathscr{D}_t = \{X_t, y_t\}$ is the training data for task t, $X_t \in \mathbf{R}^{n_T \times p_t}, y_t \in \mathbf{R}^{n_T}$, n_t is the number of samples for task t and p_t is the dimension of the samples. For each task we have a task specific function parameterized by $\theta_i \in \mathbb{R}^{d_i}$ to be estimated by a task specific loss function[2] $L_t(\theta_i; X_t, y_t)$. When learning a linear model, we typically have θ_i be the linear combination coefficients of the input features and thus $p_t = d_t$. However, when considering more complex functions, e.g., convolutional neural networks, then the parameter space is typically much larger than the input space, i.e., $d_t \gg p_t$. The multi-task learning seeks to joint optimize the loss functions from such related tasks:

$$\min_{\{\theta_i\}_{i=1}^T} \sum_{i=1}^{T} L_t(\theta_i; X_t, y_t) + \mathscr{R}\left(\{\theta_i\}_{i=1}^T\right), \tag{1}$$

where $\mathscr{R}(.)$ is a term that couples the tasks and enforces inductive knowledge transfer. The coupling term is critical due to that (a) when $\mathscr{R}(.)$ is decoupled for each task, then the learning processes are also decoupled and thus there would be no effective knowledge transfer. (b) different coupling terms convey different assumptions about how tasks are related to each other. We note that though multi-task learning focuses on the generalization performance of all tasks, we may only focus on a selected few tasks, and consider other tasks involved as *auxiliary tasks*.

In our framework, we propose to jointly learn two related tasks: a face verification task and an age estimation task, with the belief that age information is very informative when human perception of faces and that fusing in age information property could be very beneficial to the face verification task. Using aging information in cross-age face recognition problem has been previous studied in various of settings [7], [13], which has shown that improvements of

[1] In practice, the network structure could also be different, $W^F \neq W^A$

[2] The specific loss function could be cross-entropy loss of L_A shown in Eq.3 or Contrastive Loss of L_F or any other loss functions.

the face verification can be obtained by leveraging the aging information. However, though the prior work, demonstrated the relationship between face and age via decomposition, so far there is no comprehensive study on an integrated framework using end-to-end training that couples the two tasks. We use the subscript F to denote the loss function, model, data for face verification and A to denote those for the age estimation task, and we can write the multi-task objective in the following:

$$\min_{\{\theta_F,\theta_A\}} L_F(\theta_F;X_F,y_F)+\alpha L_A(\theta_A;X_A,y_A)+\mathscr{R}(\{\theta_F,\theta_A\}), \tag{2}$$

where $L_F(\theta_F;X_F,y_F)$ is the contrastive loss function that is defined as:

$$L_F(\theta_F;X_F,y_F)\triangleq\sum_{ij}Z_{ij}D_{ij}+(1-Z_{ij})\max(m-D_{ij},0), \tag{3}$$

for any pair of images (i,j) with feature distance $D_{ij}=\|\tilde{X}_i(\theta_F)-\tilde{X}_j(\theta_F)\|^2$ given model parameter θ_F, *i.e.,* $\tilde{X}_i(\theta_F)=\mathscr{P}_{\theta_F}(X_i)$, $Z_{ij}=1$ if i and j belong to the same person (a "genuine pair") and $Z_{ij}=0$ otherwise (an "impostor pair"). Positive number m acts as a margin to ensure that the engergy function for impostor pairs is larger than that of genuine pair at least by m. Clearly, The loss consists of two penalties: $Z_{ij}D_{ij}$ and $\max(m-D_{ij},0)$, while the first part penalizes a positive pair (i.e., a genuine pair) that is too far apart while the second part penalizes a negative pair (i.e., an impostor pair) that is closer than a margin m. If a negative pair is already separated by margin m, then there is no penalty for that pair.

$L_A(\theta_A;X_A,y_A)$ is the cross-entropy loss function that is defined by:

$$\begin{aligned}L_A(\theta_A;X_A,y_A)=&-\sum_{ik}y_{ik}\log[\tilde{X}_i(\theta_A)]_k\\&-\sum_{ik}(1-y_{ik})\log(1-[\tilde{X}_i(\theta_A)]_k),\end{aligned} \tag{4}$$

where $y_{ik}=1$ if the actual age[3] for image i is k and $y_{ik}=0$ otherwise; $\tilde{X}_i(\theta_A)=\mathscr{P}_{\theta_A}(X_i)$ represents the estimated age given the model/network structures θ_A with $\tilde{X}_i(\theta_A)_k=1$ if the age estimation for image i labeled k and $\tilde{X}_i(\theta_A)_k=0$ otherwise.

To summarize, L_F is the the contrastive loss function (shown in Eq. 3) defined by model parameter θ_F for face recognition while L_A is the cross-entropy loss (shown in Eq. 4) defined by model parameter θ_A for age-verification (as an auxilary task). $\alpha\in(0,1)$ is a tunable parameter balancing the two tasks.

C. Primary Task: Face Verification

There are several considerations for this framework. The first reason is due to the limited available data for cross-age face recognition. It is not easy to get comparable amount of images as collected by [37], [43]. Getting the actual age labels is even more challenging. Another consideration is that Siamese network [6] can make best use of limited data to generate enough training data. For each individual, one can generate enough positive and negative pairs. For example, for one person with n different samples, we can get $\binom{n}{2}$ positive pairs. The learning process of Siamese network can minimize a discriminative loss function which can help minimize the difference between samples of the same people and enlarge the difference between face images of different individuals due to the property of constrastive loss defined in Eq. 3.

D. Auxiliary Task: Age Estimation

Our auxiliary task of age estimation is another deep neural network that has the same architecture as the face verification task, and the only difference is that the output layer predicts age. We formulate the age estimation as a classification problem: we encode each age label into a d-dimension vector ($d=78$ in our case), each of which is an exclusive indicator for age. The cross-entropy loss is used for our age classifier as shown in Eq. 4. The age estimation task is achieved by learning a better network parameter θ_A given the training samples. In the end, one age label is used to assign the image to a particular age category that has the largest posterior probability, *i.e.,*

$$k=\arg\max_k P(y_i=k|X_i,\theta_A)=\arg\max_k\frac{\exp\left(\tilde{X}_i(\theta_A)\right)_k}{\sum_k\exp\left(\tilde{X}_i(\theta_A)\right)_k}, \tag{5}$$

where $\tilde{X}_i(\theta_A)$ represents the estimated age given the model/network structures θ_A.

E. Bridging the gap between primary and auxiliary tasks

To enable the knowledge transfer between the primary and auxiliary tasks, we take advantage of the network structures between the two tasks. We note that two networks are based on the same base architectures (VGG [38] in our case), and the difference is on the output layer. As such the knowledge transfer can be enforced on the shared network structure. Formally we explicitly write $\theta_F=\{\theta_F^S,\theta_F^I\}$, where θ_F^S includes the parameters of the shared structure and θ_F^I includes the independent parameters (parameters of the last layer). Similarly we have $\theta_A=\{\theta_A^S,\theta_A^I\}$ for the age estimation task. Based on Eq. 2, we can thus write our multi-task learning formulation as follows:

$$\begin{aligned}\min_{\theta_F=\{\theta_F^S,\theta_F^I\},\theta_A=\{\theta_A^S,\theta_A^I\}} & L_F(\theta_F;X_F,y_F)+\alpha L_A(\theta_A;X_A,y_A)\\&+\beta\|\theta_F^S-\theta_A^S\|_F^2,\end{aligned} \tag{6}$$

where $\mathscr{R}(\{\theta_F,\theta_V\})=\beta\|\theta_F^S-\theta_A^S\|_F^2$ couples the two tasks by encouraging similar values of corresponding network parameters, β specifies how much predictive information we would like to transfer from the age estimation task to the face verification task. The parameter also provides effective regularization to prevent the overfitting problem because the reduction of degree of freedom of the two networks. We note that in the extreme case $\beta\to\infty$, we would have the shared

[3]The age ranges from 1 to 78, and therefore is encoded as 78 category, where each k corresponds to an age.

TABLE I

COMPARISON OF DIFFERENT METHODS FOR CROSS-AGE FACE VERIFICATION ON MORPH. EXPERIMENTAL RESULTS ARE REFERRED FROM THE RESULTS PROVIDED BY PREVIOUS WORKS(MEASURED IN EER (%)).

Method	Year	EER (%)
Bayesian Eigenface [32]	2005	9.7
GOP [24]	2010	10.5
Bagging LDA [23]	2011	10.2
NRML [25]	2014	8.6
MNRML [25]	2014	7.5
AGCD (GOP) [7]	2015	9.2
AGCD [7]	2015	5.5
ID-DNN	Proposed	**2.7**
ID-AGE-DNN	Proposed	**2.2**

parts of two network to be identical, leading to the following:

$$\min_{\theta^S,\theta_F^I,\theta_A^I} L_F(\{\theta^S,\theta_F^I\};X_F,y_F)+\alpha L_A(\{\theta^S,\theta_A^I\};X_A,y_A). \quad (7)$$

F. Parameter Learning and Fine Tuning

To achieve very good results for parameters, we adjust the loss function defined in Eq. 7 which is actually optimized using stochastic gradient descent with momentum [19],

$$q^{t+1} \leftarrow \mu q^t - \gamma \frac{dL}{d\theta_F}(\theta_F^{t+1}),$$
$$\theta_F^{t+1} \leftarrow \theta_F^t + q^{t+1}, \quad (8)$$

where $\mu \in [0,1)$ is the momentum and $\gamma \in [0,\infty]$ is the learning rate. In our experiment, mini-batch learning algorithm is adapted where L is approximated by considering a small fraction of training samples in each iteration. θ_A is similarly learned in our experiments. Given the Siamese network that encodes the parameters θ_A, θ_F respectively, we apply the standard back-propagation algorithm to efficiently compute the gradient in Eq. 8.

Summary This framework provides a novel way to learn cross-age face verification solutions by jointly learning the framework for cross-face age verification (i.e., deciding whether two facial photos belonging to the same individual) and age estimation (to predict the actual age of the input image). The whole scheme effectively excludes distracting features in a fine-grained level for improving face verification as the main task. Furthermore, the proposed algorithm makes use of the general face feature at the low level of the deep neural network and gets the specific feature information for each task at the higher level. It successfully preserves the identity information while keeps the discriminative aging cues for the age estimation task. The detailed experimental analysis further validates the benefit of jointly face verification with age estimation tasks for improving face recognition.

IV. EXPERIMENT

A. Datasets

In this paper, we use two most popular cross-age face recognition datasets to evaluate the performance of the proposed pipeline. They are MORPH [35] and FG-NET [1] dataset. MORPH [35] is a large dataset widely used in cross-age face recognition and age estimation. MORPH includes two versions, which are MORPH Album I and MORPH Album II. Album I has very limited face images (1690 samples from 625 different people). Considering the size limitation, we use Album II as the evaluation set in the experiment. This section includes 55,132 facial images belonging to 13,617 different individuals.

Compared to MORPH, FG-NET [1] is smaller. It includes 82 different subjects. There are 1002 facial images in FG-NET. One of the most important reasons that why FG-NET is widely used in this area is because of its wide age range. In addition, it also includes these general face variations, such as expression, lighting and pose changes [30], [31]. These factors are also other common influential factors in the face verification problem. There are also low-resolution images included in the dataset.

B. Experimental Setting

For face verification, one general criteria to evaluate the performance of the system is Equal Error Rate (EER). In general, EER can be computed by calculating the value where the accept error and reject error are equal based on different thresholds using the distance between different testing samples. We use mean absolute error (MAE) to evaluate the age estimation performance. MAE is calculated based on the average of the absolute errors between the estimated age and the ground truth (labeled age). It is calculated as:

$$\text{MAE} = \frac{1}{N}\sum_{n=1}^{N}|age_n - age_n^{'}|, \quad (9)$$

where age_n represents the ground truth label of nth image, $age_n^{'}$ denotes the estimated age and N is the total number of testing samples. The smaller of MAE, better of the performance. For comparison, we also set up the experiment without adding age information where the whole network is purely a Siamese network for face verification which we name as ID-DNN. We use ID-AGE-DNN to represent the network fusing with age information. For all experiments of face verification, we use cosine-distance to measure the similarities between different facial pairs. As indicated in Fig. 1, the feature obtained in Fc_N is used as the feature descriptor.

1) MORPH Dataset: In MORPH, three-fold cross validation is applied. This setting is the same as previous works [7], [25] where 13,000 intra-individual pairs are generated by choosing image pairs of the same subject with the largest age gap. 15,000 inter-category pairs are randomly selected using images of different subjects. The whole dataset is divided into three folds. Intra-personal pairs demonstrate that image pairs belong to the same subject, whereas inter-personal pairs indicate image pairs are from different individuals. Within

TABLE II

COMPARISON WITH PREVIOUS METHODS FOR CROSS-AGE FACE VERIFICATION ON FG-NET. COMPARISON RESULTS ARE OBTAINED FROM THESE PUBLIC RESULTS AS CITED (MEASURED IN EER (%)).

Method	Year	EER (%)
Graph Matching [26]	2010	25.4
GOP [24]	2010	24.1
Landmark [49]	2010	23.6
Growth Model [49]	2012	22.3
AGCD (GOP) [7]	2015	21.7
AGCD [7]	2015	19.4
ID-DNN	Proposed	**8.0**
ID-AGE-DNN	Proposed	**7.8**

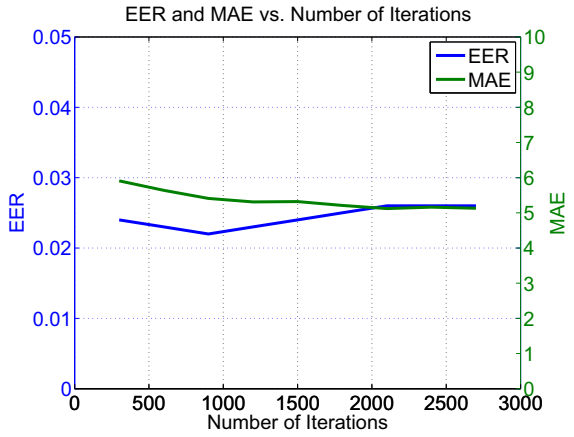


Fig. 2. Illustrations of EER and MAE variations along with different iterations .

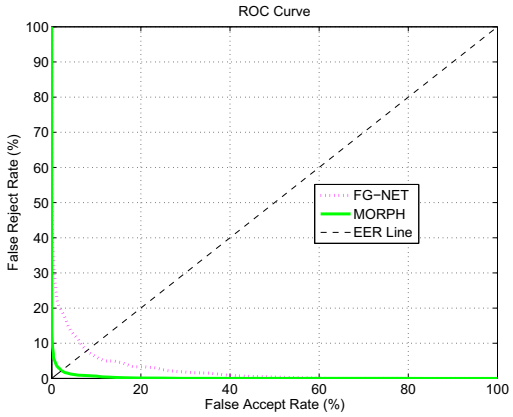


Fig. 3. Illustration of ROC curve obtained on MORPH and FG-NET using the proposed approach. Different color represents different datasets.

each fold, samples included in two folds are used for training and the left others are used for testing. There is no overlap between the training and the testing in the whole experiment.

TABLE III

COMPARISON OF MAE IN DIFFERENT AGE RANGES ON THE FG-NET DATABASE. #IMAGES[1] IS THE NUMBER OF IMAGES USED IN PREVIOUS WORK [50] FOR AGE ESTIMATION IN DIFFERENT AGE GROUPS. #IMAGES[2] IS THE NUMBER OF IMAGES USED IN THIS WORK FOR DIFFERENT AGE GROUPS.

Age	# Images[1]	# Images[2]	Proposed	CPNN [50]
10-19	339	57	**3.23**	3.83
20-29	144	105	**4.28**	8.01
30-39	70	55	**11.29**	17.91
40-49	46	37	**18.27**	25.26
50-59	15	11	**29.73**	36.40
60-69	8	7	**42.57**	45.63

The experimental results achieved on MORPH are listed in Table I. The results demonstrate that the deep learning based pipeline can improve the verification accuracy by around 3% compared to the state-of-the art. With integration of age estimation as auxiliary task, the EER goes down to 2.2%. As far as we know, this is the best verification performance achieved on MORPH dataset.

We also list EER values' changes along with different iterations in Table IV and Fig. 2. From the illustration, we can find that the best EER is obtained at around 1000 iterations. When the iteration increases further, the performance drops. One possibility is due to the reason that the size of our data is not big enough to optimize current complicated model which we are using (VGG). More iterations will cause the over-fitting problem.

2) FG-NET Dataset: Following the same protocol used in [7], [24], [26], [49], [49], we select the same adult subset of FG-NET. It contains 272 images from 62 subjects. For verification task, we generate 665 intra-personal pairs using these images. Meanwhile, we randomly select inter-personal pairs. Follow the same setting as used in [7], we generate 6000 inter-personal pairs. We use three-fold cross validation. On average, each fold includes 220 within-class pairs and 2000 between-class pairs. Within each fold, approximately 440 intra-category pairs and 4,000 extra-category pairs are used for training. With limited samples (272 samples in total), training an identity based neural network is almost impossible compared to Siamese network. During the whole experiment, there is no identity overlapping between different folds. Table III gives the performances comparison of the proposed framework to other state-of-the-arts. It shows that our algorithm achieves an ERR of 7.8%, which significantly improves over the previous best published result (19.4%) reported by [7].

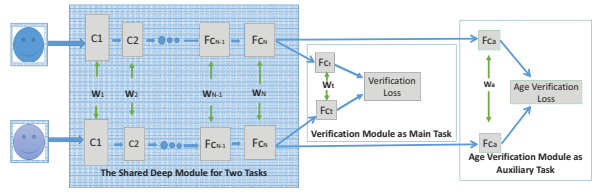
C. Implementation Details

In all our experiments, we train the CNN using Stochastic Gradient Descent (SGD) with standard backprop [44] and AdaGrad [8]. We start with a learning rate of 0.0001. To speed up the training speed, we initialized the parameters

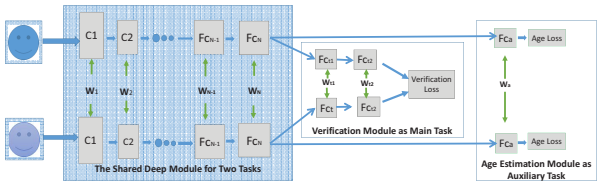
TABLE IV

ILLUSTRATION OF FACE VERIFICATION PERFORMANCE (EER) ALONG WITH AGE ESTIMATION PERFORMANCE MEAN ABSOLUTE ERROR (MAE) IN DIFFERENT ITERATIONS ON THE MORPH DATABASE.

# Iterations	300	600	900	1200	1500	1800	2100	2400	2700
MAE (ages)	5.91	5.64	5.41	5.31	5.32	5.21	5.12	5.16	5.13
EER (%)	2.4	2.3	**2.2**	2.3	2.4	2.5	2.6	2.6	2.6



(a)



(b)

Fig. 4. (a) Illustration of the comparison framework where the auxiliary task is replaced by age verification instead of age classification. (b) Illustration of the comparison framework where one additional fully-connected layer is added compared to Fig. 1.

using previous given face model [29]. All our experiments are carried using Caffe deep learning toolbox [18].

In addition, for better understanding of the proposed algorithm, we also list results using other frameworks besides our current framework. Instead of building age estimation sub-tasks, we also build the age verification network as proposed in [7] (to decide if two given face images have the same age or not) and fuse it into the verification framework.

D. Comparison between Different Schemes

The main motivation in this study is to improve cross-age face verification performance with age information. The general illustration of this schemes is listed in Fig. 4. It includes two comparison schemes. The details of these schemes are illustrated as follows:

Fig. 4(a) illustrates the pipeline where we put the age estimation as verification task talked in [7]. There are two sets of labels used in this framework. One set of the label is used to indicate if two facial images belong to the same individual or not. The other label set is used to indicate if two input images have the same age or not. We use 1 to indicate that they are same age and 0 to represent different ages. Using this scheme, our experimental results demonstrate that the best EER obtained on MORPH is more than 5%. This shows that using age verification as auxiliary task is inferior to using age classification task for improving the face verification performance.

The scheme listed in Fig. 4(b) is the framework where we add another fully-connected layer Fc_{t+1} before the original full-connected layer Fc_t for calculating the contrastive loss for face verification. We have tested the performance using Fc_t as the feature vector for verification. The verification performance is pretty low where EER is greater than 7% on MORPH dataset. Besides these two illustrated frameworks, we also tried to increase the layer in depth both for age and verification path. However, this does not help improve the verification performance.

E. Age Estimation

Besides face verification task, we also evaluate the age estimation performance as listed in Table III and Table IV. The trade-off between identity and age task can be set via the loss weight. In our work, we set the loss weights to specify their relative importance. By increasing the relative loss weight of age estimation auxiliary task to the identity task, in this study, the best MAE on MORPH we can get is 3.01. If we increase the loss weight for the identity task, MAE increases with the iterations go up as listed in Table IV. Although previous works show that age and identity division can help improve face verification performance. However, these works did not discuss the performance of the feature for age estimation task. As far as we know, our study first provides the actual age estimation result along with face verification task in one joint framework.

V. CONCLUSION

In this paper, we have proposed a novel multi-task deep neural network architecture for cross-age face verification. Our approach sets face verification as the primary learning task and age estimation as the auxiliary learning task. By exploiting the intrinsic, shared low-dimensional representation, the proposed algorithm can effectively balance feature sharing and feature exclusion between the two tasks. Experimental results show that the proposed method significantly outperforms the state-of-the-art by a large margin. Moreover, the proposed multi-tasking learning scheme demonstrates that integrating age information via a unified deep neural network can boost the performance of cross-age face verification. Our future work includes further evaluating the proposed framework using more datasets and exploring more discriminative loss functions.

REFERENCES

[1] The fg-net aging database `http://www.fgnet.rsunit.com/`.
[2] H. Chang, Y. Zhou, P. Spellman, and B. Parvin. Stacked predictive sparse coding for classification of distinct regions in tumor histopathology. In *ICCV*, 2013.

[3] K.-Y. Chang, C.-S. Chen, and Y.-P. Hung. Ordinal hyperplanes ranker with cost sensitivities for age estimation. In *CVPR*, pages 585–592, 2011.

[4] W.-L. Chao, J.-Z. Liu, and J.-J. Ding. Facial age estimation based on label-sensitive learning and age-oriented regression. *Pattern Recognition*, 46(3):628–641, 2013.

[5] B.-C. Chen, C.-S. Chen, and W. H. Hsu. Cross-age reference coding for age-invariant face recognition and retrieval. In *ECCV 2014*, pages 768–783. Springer, 2014.

[6] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *CVPR*, volume 1, pages 539–546. IEEE, 2005.

[7] L. Du and H. Ling. Cross-age face verification by coordinating with cross-face age verification. In *CVPR*, pages 2329–2338, 2015.

[8] J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning Research*, 12:2121–2159, 2011.

[9] Y. Fu and T. S. Huang. Human age estimation with regression on discriminative aging manifold. *IEEE Transactions on Multimedia*, 10(4):578–584, 2008.

[10] X. Geng, C. Yin, and Z.-H. Zhou. Facial age estimation by learning from label distributions. *IEEE Transactions on Pattern Analysis and Machine Intelligence,*, 35(10):2401–2412, 2013.

[11] X. Geng, Z.-H. Zhou, and K. Smith-Miles. Automatic age estimation based on facial aging patterns. *PAMI*, 29(12):2234–2240, 2007.

[12] D. Gong, Z. Li, D. Lin, J. Liu, and X. Tang. Hidden factor analysis for age invariant face recognition. In *ICCV*, pages 2872–2879, 2013.

[13] D. Gong, Z. Li, D. Tao, J. Liu, and X. Li. A maximum entropy feature descriptor for age invariant face recognition. In *CVPR*, pages 5289–5297, 2015.

[14] P. Gong, J. Zhou, W. Fan, and J. Ye. Efficient multi-task feature learning with calibration. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 761–770. ACM, 2014.

[15] G. Guo, G. Mu, Y. Fu, and T. S. Huang. Human age estimation using bio-inspired features. In *CVPR*, pages 112–119, 2009.

[16] G. Guo and X. Wang. A study on human age estimation under facial expression changes. In *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2547–2553. IEEE, 2012.

[17] H. Han, C. Otto, X. Liu, and A. Jain. Demographic estimation from face images: Human vs. machine performance. *PAMI*, 2014.

[18] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the ACM International Conference on Multimedia*, pages 675–678. ACM, 2014.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[20] Y. H. Kwon and N. da Vitoria Lobo. Age classification from facial images. In *CVPR*, pages 762–767, 1994.

[21] G. Levi and T. Hassner. Age and gender classification using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 34–42, 2015.

[22] C. Li, Q. Liu, J. Liu, and H. Lu. Learning ordinal discriminative features for age estimation. In *CVPR*, pages 2570–2577, 2012.

[23] Z. Li, U. Park, and A. K. Jain. A discriminative model for age invariant face recognition. *IEEE Transactions on Information Forensics and Security*, 6(3):1028–1037, 2011.

[24] H. Ling, S. Soatto, N. Ramanathan, and D. W. Jacobs. Face verification across age progression using discriminative methods. *IEEE Transactions on Information Forensics and Security*, 5(1):82–91, 2010.

[25] J. Lu, X. Zhou, Y.-P. Tan, Y. Shang, and J. Zhou. Neighborhood repulsed metric learning for kinship verification. *PAMI,*, 36(2):331–345, 2014.

[26] G. Mahalingam and C. Kambhamettu. Age invariant face recognition using graph matching. In *2010 Fourth IEEE International Conference on Biometrics: Theory Applications and Systems (BTAS),*, pages 1–7. IEEE, 2010.

[27] B. Ni, Z. Song, and S. Yan. Web image and video mining towards universal and robust age estimator. *IEEE Transactions on Multimedia*, 13(6):1217–1229, 2011.

[28] U. Park, Y. Tong, and A. K. Jain. Age-invariant face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence,*, 32(5):947–954, 2010.

[29] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *BMVC*, 2015.

[30] X. Peng, X. Yu, K. Sohn, D. Metaxas, and M. Chandraker. Reconstruction for feature disentanglement in pose-invariant face recognition. *arXiv preprint arXiv:1702.03041*, 2017.

[31] X. Peng, S. Zhang, Y. Yang, and D. N. Metaxas. Piefa: Personalized incremental and ensemble face alignment. In *ICCV*, 2015.

[32] N. Ramanathan and R. Chellappa. Face verification across age progression. *IEEE Transactions on Image Processing,*, 15(11):3349–3361, 2006.

[33] N. Ramanathan and R. Chellappa. Modeling age progression in young faces. In *CVPR*, volume 1, pages 387–394, 2006.

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

[35] K. Ricanek and T. Tesafaye. Morph: A longitudinal image database of normal adult age-progression. In *FG*, pages 341–345, 2006.

[36] M. Riesenhuber and T. Poggio. Hierarchical models of object recognition in cortex. *Nature neuroscience*, 2(11):1019–1025, 1999.

[37] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, pages 815–823, 2015.

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

[39] Y. Sun, X. Wang, and X. Tang. Deep learning face representation from predicting 10,000 classes. In *CVPR*, pages 1891–1898, 2014.

[40] Y. Sun, X. Wang, and X. Tang. Deeply learned face representations are sparse, selective, and robust. In *CVPR*, pages 2892–2900, 2015.

[41] D. Sungatullina, J. Lu, G. Wang, and P. Moulin. Multiview discriminative learning for age-invariant face recognition. In *FG*, pages 1–6. IEEE, 2013.

[42] J. Suo, S.-C. Zhu, S. Shan, and X. Chen. A compositional and dynamic model for face aging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(3):385–401, 2010.

[43] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*, pages 1701–1708, 2014.

[44] Y. Tsuruoka, J. Tsujii, and S. Ananiadou. Stochastic gradient descent training for l1-regularized log-linear models with cumulative penalty. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 477–485. Association for Computational Linguistics, 2009.

[45] X. Wang, R. Guo, and C. Kambhamettu. Deeply-learned feature for age estimation. In *2015 IEEE Winter Conference on Applications of Computer Vision (WACV),*, pages 534–541. IEEE, 2015.

[46] X. Wang and C. Kambhamettu. Age estimation via unsupervised neural networks. In *FG*, 2015.

[47] X. Wang, R. Li, Y. Zhou, and C. Kambhamettu. A study of convolutional sparse feature learning for human age estimation. In *FG*, 2017.

[48] X. Wang, V. Ly, G. Lu, and C. Kambhamettu. Can we minimize the influence due to gender and race in age estimation? In *ICMLA*, volume 2, pages 309–314, 2013.

[49] T. Wu, P. Turaga, and R. Chellappa. Age estimation and face verification across aging using landmarks. 2012.

[50] C. Y. Xin Geng and Z.-H. Zhou. Facial age estimation by learning from label distributions. *PAMI*, 2013.

[51] D. Yi, Z. Lei, and S. Z. Li. Age estimation by multi-scale convolutional network. In *ACCV 2014*, pages 144–158. Springer, 2014.

[52] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.

[53] T. Zhang, B. Ghanem, S. Liu, and N. Ahuja. Robust visual tracking via structured multi-task sparse learning. *International journal of computer vision*, 101(2):367–383, 2013.

[54] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Facial landmark detection by deep multi-task learning. In *ECCV 2014*, pages 94–108. Springer, 2014.

[55] J. Zhou, J. Chen, and J. Ye. Clustered multi-task learning via alternating structure optimization. In *Advances in neural information processing systems*, pages 702–710, 2011.

[56] J. Zhou, J. Chen, and J. Ye. Malsar: Multi-task learning via structural regularization. *Arizona State University*, 21, 2011.

[57] J. Zhou, L. Yuan, J. Liu, and J. Ye. A multi-task learning formulation for predicting disease progression. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 814–822. ACM, 2011.