Deep Aging Face Verification With Large Gaps

Luoqi Liu, Chao Xiong, Hanwang Zhang, Zhiheng Niu, Meng Wang, and Shuicheng Yan, *Senior Member, IEEE*

***Abstract*—Along with the long-time evolution of popular social networks, e.g. Facebook, social media analysis research inevitably arrived at the era of considering face/user recognition with large age gaps. However, related research with adequate subjects and large age gaps is surprisingly rare. In this work, we first collect a so-called cross-age face (CAFE) dataset, ranging from child, to young, to adult, to old groups. Then, we propose a novel framework, called deep aging face verification (DAFV), for this challenging task. DAFV includes two modules: aging pattern synthesis and aging face verification. The aging pattern synthesis module synthesizes the faces of all age groups for the input face of an arbitrary age, and the core structure is a deep aging-aware denoising auto-encoder (a^2-DAE) with multiple outputs. The aging face verification module then takes the synthesized aging patterns of a face pair as the input, and each pair of synthesized images of the same age group is fed into a parallel CNN; finally, all parallel CNN outputs are fused to provide similar/dissimilar prediction. For DAFV, the training of the aging face verification module easily suffers from the overfitting results from the aging pattern synthesis module, and we propose to use the cross-validation strategy to produce error-aware outputs for the synthesis module. Extensive experiments on the CAFE dataset well demonstrate the superiority of the proposed DAFV framework over other solutions for aging face verification.**

***Index Terms*—Cross-age, deep learning, face verification.**

I. Introduction

With the growing popularity of digital devices, it has been increasingly convenient for people to share photos on various websites, such as Facebook and Flickr. These photos function as a way of connection with other people, and many of them may contain human faces. These considerable number of face images provide rich research material for multimedia studies and benefit many valuable applications, such as face recognition [18], similar face matching, face annotation [10] and face retrieval [8]. To date, many photo sharing websites have been working for quite a long time, and may continue to provide services in an expected long term. The recognition of the user's faces with large age gaps has become a great challenge for most existing applications. When a person grows old, his/her face appearance may change a lot, which makes it difficult to correctly recognize his/her photos captured at early ages from the photo album or social network. To address this problem, we build a novel system as shown in Fig. 1, which can recognize face images across large age gaps. Given a pair of face images from possibly different age groups as the input, our system will synthesize faces of all the age groups for each of them. Then, an aging face verification process is followed to recognize whether the input pair is from the same identity. Since the two steps are both built on deep learning [16], [21], we call the framework as Deep Aging Face Verification (DAFV).

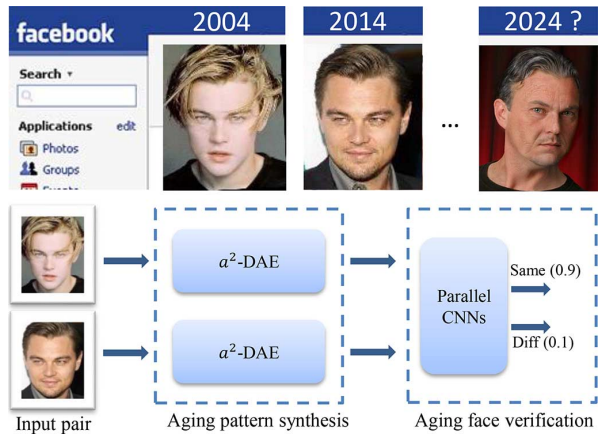


Fig. 1. Illustration of our two-stage deep learning system for aging face verification with large gaps. Given a face pair as the input, our system will synthesize the faces of all the age groups, and then verify whether they belong to the same person.

The difficulty in developing such a system mainly lies in the facial appearance change during the aging process. Here, the facial appearance includes the facial shape and texture. Facial aging is a very complex process, which involves changes in both the facial shape and texture. As a person grows from young to old, the facial shape will alter as the skull grows, and the facial texture will also show wrinkles gradually. Without taking the facial appearance change across ages into consideration, the performance of current face recognition systems may be degraded.

Despite its practical significance, face recognition with large age gaps is a very challenging problem which has rarely been studied. Till now there has been no satisfactory dataset for the research on this problem in multimedia society. Several widely used databases for cross-age face recognition research, i.e., FG-NET [12], MORPH [28] and CACD [7], all have limitations. FG-NET [12] only contains 82 subjects, which is limited in diversity in terms of human race, gender, living environment, etc. CACD dataset contains more than 160 K images of 2 K celebrities, which is by far the largest cross-age face dataset. But the age gap of the photos from the same person in this dataset is quite small, which is not suitable to study the long

Manuscript received April 14, 2015; revised August 28, 2015 and October 20, 2015; accepted November 12, 2015. Date of publication November 13, 2015; date of current version December 14, 2015. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Enrico Magli.

L. Liu, Z. Niu, and S. Yan are with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117583.

H. Zhang is with the School of Computing, National University of Singapore, Singapore 117417.

C. Xiong is with Department of Electrical and Electronic Engineering, Imperial College London, London W3 0DF, U.K.

M. Wang is with the School of Computer and Information, Hefei University of Technology, Hefei 230009, China.

Digital Object Identifier 10.1109/TMM.2015.2500730

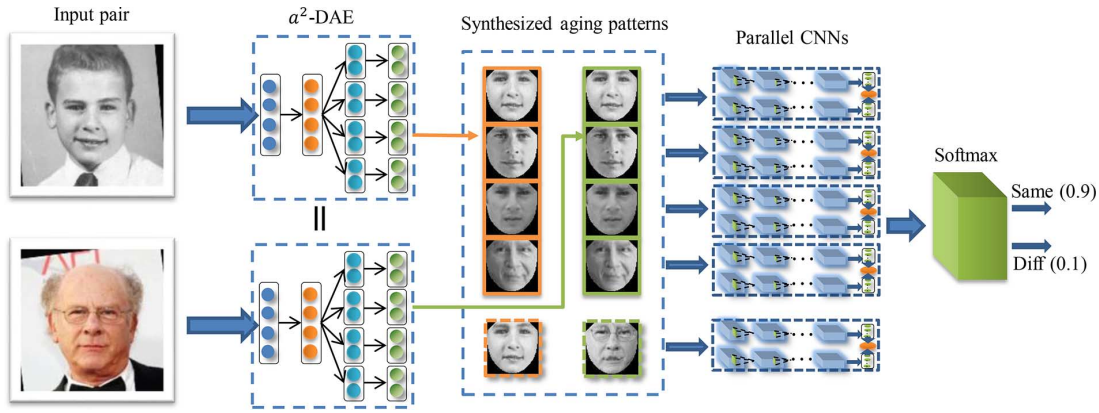


Fig. 2. Flowchart of the DAFV architecture. It includes two modules: the aging pattern synthesis module and the aging face verification module. In the aging pattern synthesis module, a novel aging-aware DAE (a^2-DAE) is proposed to synthesize the faces of all the age groups. In the aging face verification module, parallel CNNs are trained based on the synthesized faces and the original faces to predict the verification score.

time span aging process. MORPH [28] faces the similar small age gap issue. Therefore, a new database which can cover both large human identities and large age gaps is desired in this research area. Besides, existing research work, which is not much, mainly tackles the cross-age face recognition problem from either of two directions. The first group of researchers, such as Li *et al.* [24], focus on designing new features, which possess a high discriminative ability to recognize different people and robustness against age changes. The other group of researchers [27], [37] model the face appearance change caused by face aging, but do not well explore the discriminative features.

Based on these observations, we address the cross-age face verification problem from two perspectives. Firstly, we build the Cross-Age FacE (CAFE) dataset, which contains 901 male and female celebrities. The collection of this dataset is not trivial work. Most identities only have photos of a narrow age range. We initially create a list of ∼10 K person names, and only 901 identities are remained after filtering. Compared with previous cross-age face datasets, the CAFE dataset achieves a good balance between identities coverage and age range. People included in CAFE are of various races and nationalities, which largely increases the diversity of the proposed dataset. Moreover, the large range of ages for each subject renders CAFE a suitable dataset for modeling the complex aging processes across ages.

Secondly, we propose a new deep learning architecture, called Deep Aging Face Verification (DAFV), to model the aging process and then extract strong discriminative features for the verification task. The flowchart of this architecture is illustrated in Fig. 2. Two modules are included in this architecture: the aging pattern synthesis module and the aging face verification module. The aging process is well modeled in the aging pattern synthesis module. Without knowing groundtruth age of the input face, this module can synthesize faces in all age ranges. The synthesized faces can be visualized and make aging process easily understood. The later verification module can also benefit from this module. We can directly compare face pairs in the same age range to eliminate face appearance change caused by aging. Based on the synthesized faces in all the age ranges, verification performance can be improved. We use a novel deep aging-aware denoising auto-encoder (a^2-DAE) to synthesize the face appearances of all the age groups for the input face of an arbitrary age. Different from the ordinary DAE, the aging-aware a^2-DAE has multiple branches in the decoding layer. The reconstructed face of a certain age group can be output from the corresponding branch. Then the faces for all the age groups are synthesized for an input image pair to the verification module, the following verification can be conducted by directly comparing the faces from the same age group. In this way, the age difference of a face pair is well dismissed.

For the aging face verification module, with the synthesized aging patterns of the face pair taken as the input, each pair of synthesized faces from the same age group is fed into a parallel convolutional neural network (CNN) [21] to learn discriminative features and do the face verification. Each CNN model only focus a single age range, and all the CNN models are fused to achieve better performance. Unlike the traditional methods, the convolutional neural network can directly learn strong discriminative features supervised by the pair labels. In the parallel CNN, a discriminative space is learned where the similarity of the face pair from the same identity is maximized, while that of the pair from different identities is minimized. These parallel CNNs trained based on the synthesized faces of each age group are then jointly fine-tuned to achieve a high discriminative capacity.

However, the DAFV architecture can easily suffer from overfitting in the aging pattern synthesis module. In the aging pattern synthesis module, the square error loss between groundtruth aging pattern and the reconstructed aging pattern is used as objective function of the a^2-DAEs. Because of overfitting, the objective function values can be low in the training set, while those are generally higher in the testing set. This will lead to "perfect" reconstruction of the aging patterns in the training set, but will cause large reconstruction errors in the testing set. This mis-match will result in the bad generalization capability of the trained parallel CNNs which are based on the synthesized faces. To avoid this problem, we train the aging pattern synthesis module in a cross-validation fashion, such that the training and testing reconstruction errors are well balanced. Then, the synthesized aging patterns are fed into the aging face verification module, which have the similar reconstruction errors from the training and testing sets.

The contributions of this work are summarized as follows.

- A large Cross-Age FacE (CAFE) dataset is constructed, including 4 650 face images of 901 celebrities covering large age gaps, which can serve as a new and comprehensive benchmark for the research community to study the aging face verification problem.
- The Deep Aging Face Verification (DAFV) architecture is proposed, including two modules, i.e., aging pattern synthesis module and aging face verification module.
- A novel training strategy is exploited to produce error-aware outputs based on the cross-validation strategy for the aging synthesis module, such that the whole framework can less suffer from overfitting.

The remaining sections are organized as follows. Section II reviews the related work. Section III introduces the CAFE dataset. Section IV describes the whole framework, including the aging pattern synthesis module and the aging face verification module. Experiments are presented in Section V. Finally, Section VI concludes our work and discusses future work.

II. Related Work

There have been very few datasets for the cross-age face recognition research. FG-NET [12], MORPH [28] and CACD [7] databases are the most widely used face databases, which serve as evaluation benchmarks for cross-age face recognition methods [27], [24], [37]. The FG-NET database contains only 1 002 images of 82 subjects from age 0 to 69. The relatively small size of the database makes it inappropriate for the real applications. The MORPH database contains two subsets: MORPH album 1 and MORPH album 2. MORPH album 1 contains 1 690 images of 625 subjects, and MORPH album 2 contains 15 204 images of 4 039 subjects. Cross-Age Celebrity Dataset (CACD) is the largest publicly available cross-age face dataset. It contains more than 160 K images of 2K celebrities with age ranging from 16 to 62. The photos in this dataset are collected from the Internet within the range of ten years. However, each subject in the MORPH and CACD database only has images with a small age gap, which makes it inappropriate for modeling the aging process for aging face recognition with large age gaps.

The research literature on cross-age face recognition is also quite limited over the past decades. Geng *et al.* [12] modeled the face aging patterns. The face aging pattern is defined as a sequence of face images from the same person sorted in the time order. A principal component space of aging pattens is constructed to model the correlation of faces from different age groups. The faces at different ages of the testing face can be reconstructed by projecting the testing face into the subspace. Park *et al.* [27] proposed a 3D aging model, which can capture the aging pattern in the 3D domain. They first converted 2D images into 3D ones by a 3D morphable model [6], and then the facial shape and texture changes are modeled separately in the Principal Component Analysis (PCA) [20] subspace. The missing samples in the training set will be generated by interpolating from the samples of the nearest ages. Wu *et al.* [35] used a parametric craniofacial growth model to model the facial shape change. These methods can model the aging process of the face shape or texture, but are weak in the discriminative capacity. Li *et al.* [24] proposed a discriminative model for age-invariant face recognition. They used scale invariant feature transform (SIFT) and multi-scale local binary patterns (MLBP) as local descriptors. To avoid overfitting, multi-feature discriminant analysis (MFDA) was proposed to process the two local feature spaces in a unified framework. It focused on highly discriminative features but failed to model the aging process. With the help of deep learning methods, our proposed framework can not only model and synthesize the aging process, but also learn discriminative features to achieve high performance.

There have been many works exploiting deep learning technology for face analysis/recognition problem. Based on deep belief networks, Luo *et al.* [25] propose a novel face parse, which can hierarchically parse faces into parts, components and pixel-wise labels. Taigman *et al.* [33] and Sun *et al.* [30], [29] use convolutional neural networks (CNN) [21] based methods for face verification problem. The performance of their works already reaches or surpasses human's performance on the widely used labeled face in the wild (LFW) dataset [19]. Zhu *et al.* [36] propose a novel multi-view perception network (MVP), which can reconstruct a full spectrum of views based on a single 2D face. However, all these works have not taken cross-age face verification problem into consideration, which is the main problem we want to handle in our work.

TABLE I
Statistics of Age Groups

Age Groups	#Faces	#Age Ranges
Child	527	$0 \sim 12$
Young	2177	$13 \sim 25$
Adult	1612	$26 \sim 50$
Old age	343	>50

III. The Cross-Age Face (CAFE) Dataset

The Cross-Age FacE (CAFE) dataset is constructed with photos of 901 celebrities. We first collect a list of $\sim$10 K celebrities' names and crawl images from the Internet. These celebrities include actors, singers and politicians. The genders are roughly balanced. Then face bounding boxes and 68 landmark points in the faces are detected and located by OMRON facial analysis toolbox.[1] Faces are aligned by similarity transform according to the centers of two eyes and that of the mouth. The distance of two eye centers is set to 32 pixels. The photos are cropped with enlarged face bounding boxes of size 160 $\times$ 160, and then saved. The images which contain non-frontal faces are removed, and the remaining photos have only near frontal faces. We finally collect 4 659 photos of 901 celebrities, and each identity has $\sim$5 faces on average.

Based on the photos' taken date stored in the metadata and the celebrities' years of birth, we divide the photos into four age groups. The identity and age labels are manually examined to guarantee label correctness. The dividing criterion and the number of faces in each age group are listed in Table I. Some celebrities have photos of all the four age groups, while some have photos of two or three age groups. Our dataset contains more subjects than the FG-NET dataset and has much larger age gaps for each subject than the MORPH dataset. Some exemplar

[1]"OMRON, OKAO Vision," [Online]. Available: http://www.omron.com/r_d/coretech/vision/okao.html

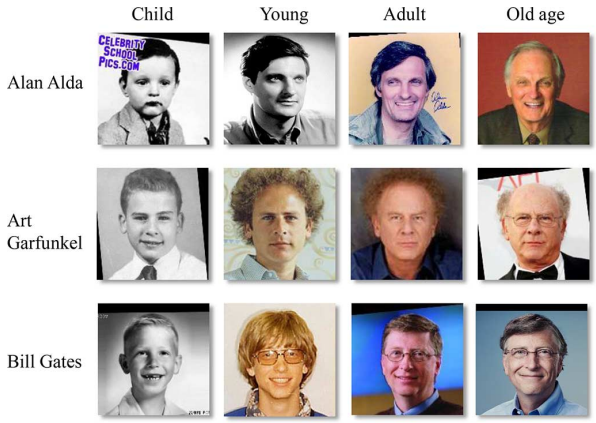


Fig. 3. Some exemplar faces in the CAFE dataset. From top to bottom, the celebrities are Alan Alda, Art Garfunkel, and Bill Gates. The photos are shown in four age groups: child, young, adult, and old age, from the first column to the last column.

faces of the celebrities from the CAFE dataset are shown in Fig. 3.

IV. Deep Aging Face Verification

A. Framework Overview

Our proposed whole framework for cross-age face verification includes the following steps.

- *Preprocessing: shape and texture separation*. Faces are preprocessed to extract shape and shape-free texture, as described in Section IV-B.
- *Aging pattern synthesis module*. The deep aging-aware denoising auto-encoder (a^2-DAE) is learned to synthesize the faces at all the age groups for the input face. The details will be described in Section IV-C.
- *Aging face verification module*. Given aging pattern pair as input, the parallel convolutional neural network is exploited to learn a discriminative space for the verification task.

B. Preprocessing: Shape and Texture Separation

Both shape and texture of a face contain important information about human age and identity. The cranial size of a face increases quickly as a person grows until 19 years old. After that, the facial texture change becomes the dominant factor for human aging [2]. Wrinkles are deepened at the sides of the eyes, and freckles and aging spots occur on the face skin. However, shape and texture correlate with each other deeply on the face, and are also influenced by other factors, such as pose and illumination. This phenomenon makes cross-age face verification an even more challenging problem.

Based on the above observations, we extract shape and texture from the faces and model them separately. 68 face landmark points are located by the OMRON face alignment algorithm. Faces are aligned according to the centers of two eyes and that of the mouth. The shape information is represented by the normalized coordinates of landmark points on the aligned faces.

To extract shape-free texture, we first calculate the mean shape of all the training images from the dataset. Delaunay triangulation [11] is computed on the mean face and each face image in the dataset to obtain 111 triangles. Piecewise linear affine transformation [14] is applied within the corresponding triangles between the face image and the mean face to obtain the warped face. Fig. 4 shows the shape-free texture extraction process. Fig. 4(a) is the original face image. Fig. 4(b) shows the detected 68 landmark points. The extracted 111 triangles are shown in Fig. 4(c). Fig. 4(d) illustrates the warped face after piecewise linear affine transformation, which represents the shape-free texture.

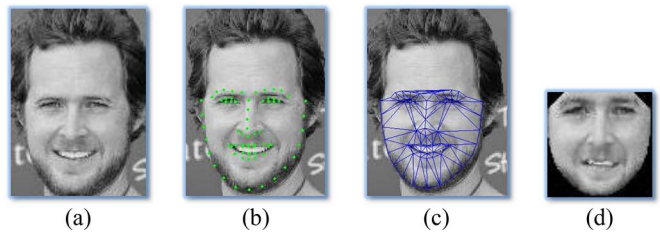


Fig. 4. Example of the shape-free texture extraction process: (a) the original face, (b) the detected face landmark points, (c) the Delaunay triangulation, and (d) the warped face.

C. Aging Pattern Synthesis Module

1) Motivations: Facial appearance (shape and texture) changes dramatically along with the human aging process, which poses a great challenge to current face verification systems. For example, if we directly compare two face images of the same person, one for the childhood and the other for the adult, due to the changes in face shape and texture over such a large time span caused by the environment, genes, and other social factors, the similarity between the two faces in the feature space may be low. Most current face recognition systems may fail in such a case. Thus, modeling the face appearance change over the time is a necessary step for cross-age face recognition.

Unlike other factors such as gender or facial expression, face aging has its own characteristics. First of all, human aging is personalized. One's face appearance is determined by mainly two aspects: internal factors, i.e., genes, and external factors, such as one's living environment, lifestyle, etc. Genes determine the initial appearance of a person. As the person grows up, many external factors may impose their influence on what he/she looks like. For example, a man who has an unhealthy diet tends to have a fat face. Secondly, face aging is an irreversible sequential process. Every person, if no deathly accident or disease occurs, experiences the growing process from childhood, youth to adult and old age, in a temporal order. No one can go through the process the other way around. It is slow with decades of time, but irreversible.

Based on the characteristics, the face appearance change of each subject should be considered as a function of both identity and age. Each image I in the cross-age face dataset should have two labels: the identity label $id(I)$ and the age label $age(I)$. This is the difference between the ordinary face verification problem and the cross-age face verification problem. For the ordinary face verification problem, given two input faces I_a and I_b, the system verifies whether $id(I_a)$ equals $id(I_b)$. No age information is considered in the ordinary face system. In the cross-age face verification system, given two input faces I_a at $age(I_a) \in \{child,\ young,\ adult,\ old\ age\}$ and I_b at $age(I_b) \in \{child,\ young,\ adult,\ old\ age\}$, the system verifies whether $id(I_a)$ equals $id(I_b)$, but does not require $age(I_a) =$

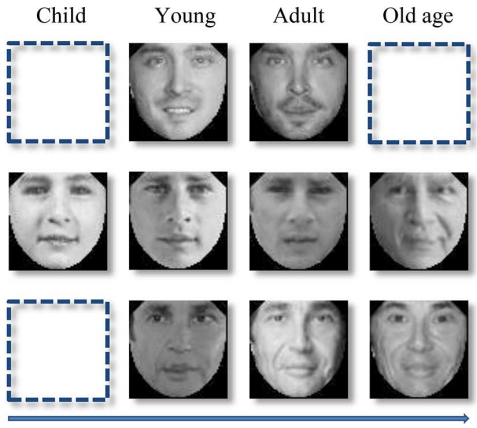


Fig. 5. Examples of the aging pattern. Each row from left to right is the aging pattern sorted in the time order. The blank bounding box shows the missing position in the aging pattern.

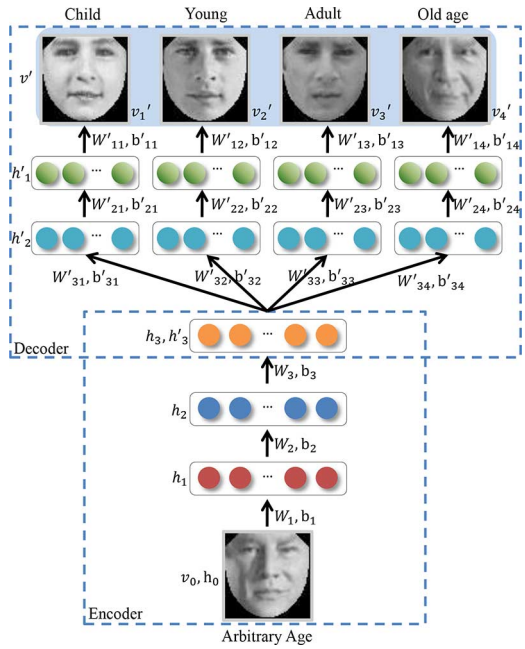


Fig. 6. Deep aging-aware denoising auto-encoder (a^2-DAE).

$age(I_b)$. If we can map I_a and I_b to the same age group as $\hat{I}_a$ and $\hat{I}_b$ where $age(\hat{I}_a) = age(\hat{I}_b)$, and directly verify whether $id(\hat{I}_a)$ equals $id(\hat{I}_b)$, the face verification problem will be much easier.

We follow Geng *et al.* [12] to represent the face appearance change of the same person over the time as the *aging pattern*. An aging pattern is defined a sequence of face images of the same identity sorted in the temporal order [12]. The aging pattern is personalized and ordered by time. Here, we consider four time spans: child, young, adult and old age. Fig. 5 shows some examples of aging patterns. The aging pattern of one person is shown in one row. From the first to the last column, the aging pattern is sorted in the time order. Note that some positions in the aging patterns of the training samples are missing, so the aging model should have the ability to handle the missing aging patterns. In the testing stage, the aging pattern of the testing sample will be synthesized. Based on the synthesized aging pattern, the testing pair I_a and I_b are projected to $\hat{I}_a$ and $\hat{I}_b$, where $age(\hat{I}_a) = age(\hat{I}_b)$. Thus, the facial appearance change caused by aging is well eliminated.

2) The Deep Aging-Aware Denoising Auto-Encoder: We propose the deep aging-aware denoising auto-encoder (a^2-DAE) to learn the aging model. Using the aging patten of each face image as groundtruth, the a^2-DAE is trained in a supervised way. Given a testing image as the input, the aging model will predict the aging pattern of the testing image. We first review some basic concepts of the auto-encoder, and then go ahead to our formulation of the a^2-DAE.

Auto-encoder and denoising auto-encoder: Given an image as the input, in which the pixel values are considered as the visible variables $\mathbf{v}$, the auto-encoder [4] first encodes it into a hidden representation $\mathbf{h}$ via a deterministic mapping $\mathbf{h} = \sigma(\mathbf{W}\mathbf{v} + \mathbf{b})$, where σ is the activation function, such as the *sigmoid* function. The hidden representation $\mathbf{h}$ is then decoded back into $\mathbf{v}'$, the prediction of $\mathbf{v}$, through a similar mapping function $\mathbf{v}' = \sigma(\mathbf{W}'\mathbf{h}+\mathbf{b}')$. If the reverse weight parameter $\mathbf{W}'$ is constrained to $\mathbf{W}' = \mathbf{W}^{\mathbf{T}}$, then $\mathbf{W}'$ is called the *tied weight*. The reconstruction error can be measured in many ways, such as cross-entropy and squared loss. We use squared loss in our later formulation as the loss function, where $L(\mathbf{v}, \mathbf{v}') = \|\mathbf{v} - \mathbf{v}'\|^2$. The hidden representation $\mathbf{h}$ is viewed as a lossy compressed code of $\mathbf{v}$. If there is only one linear hidden layer and the squared loss is used as the cost function, the $k_{\mathbf{h}}$ hidden units of $\mathbf{h}$ can be viewed as the first $k_{\mathbf{h}}$ principal components of the training data. Then the auto-encoder can be viewed the same as PCA. When the nonlinear activation function σ, such as *sigmoid* or *tanh* is used, the auto-encoder will behave differently from PCA. To force the hidden layer $\mathbf{h}$ recovering more robust features and prevent it simply overfitting to the training data, the input data can be stochastically corrupted by noises. This corrupted version of auto-encoder is called *denoising auto-encoder*.

Stacked denoising auto-encoder: The denoising auto-encoder can be stacked into multiple layers to form the stacked denoising auto-encoder. This network should firstly be pretrained layer by layer in an unsupervised manner. After the k-th layer is pretrained, then taking the hidden representation $\mathbf{h}_k$ of the k-th layer as the input, the $(k+1)$-th layer will be pretrained. The pretraining can be conducted by the auto-encoder or Restricted Boltzmann Machine (RBM) [4]. After the pretraining stage, a supervised fine-tuning process can be conducted to jointly train the whole model from the first layer to the last layer.

The deep aging-aware denoising auto-encoder (a^2-DAE) is shown in Fig. 6. Denote the input image in the first layer as $\mathbf{v}_0 \in \mathbb{R}^d$ (or $\mathbf{h}_0$ for convenience) and the aging pattern as $\{\mathbf{v}_i \in \mathbb{R}^d | i = 1 \dots 4\}$. The a^2-DAE aims to learn a mapping function $f(\mathbf{v}_0)$ to give the reconstruction of the aging pattern as $\{\mathbf{v'}_i \in \mathbb{R}^d | i = 1 \dots 4\}$. $d = 64 \times 64 = 4\,096$ is the dimension of the training and testing images. The mapping function $f(\mathbf{v}_0)$ can be decomposed as follows:

$$\begin{aligned} \mathbf{h}_i &= \mathbf{W}_i\sigma(\mathbf{h}_{i-1}) + \mathbf{b}_i, \quad i = 1, 2, 3 \\ \mathbf{h}'_{kj} &= \mathbf{W}'_{k+1,j}\sigma(\mathbf{h}'_{k+1}) + \mathbf{b}'_{k+1,j}, \quad k = 2, 1, j = 1 \dots 4 \\ \mathbf{v}'_j &= \mathbf{W}'_{1j}\sigma(\mathbf{h}'_{1j}) + \mathbf{b}'_{1j}, \quad j = 1 \dots 4. \end{aligned} \tag{1}$$

We use the *sigmoid* function as the activation function $\sigma(h) = (1 + \exp(-h))^{-1}$. $\mathbf{W}$ and $\mathbf{b}$ are the weight matrices and bias vectors. $\mathbf{h}_i, i = 1, 2, 3$ is the hidden representation of the input

data. The numbers of hidden units of $\mathbf{h}_1, \mathbf{h}_2, \mathbf{h}_3$ are 2 500, 1 000 and 400, respectively. $\mathbf{h}'_{kj}, k = 2, 1, j = 1 \ldots 4$ is the reconstructed hidden representation $\mathbf{h}_k$ at position j. $\mathbf{h}'_{kj}$ has the same number of hidden units with $\mathbf{h}_k$. $\mathbf{v}'_j$ is the reconstructed aging pattern at position j. $\mathbf{h}_3$ is a shared representation across all the modalities, which constructs an age-invariant space for all the input images with different age labels.

To train the a^2-DAE, we minimize the square error loss between ground-truth and reconstruction of the aging pattern

$$\min_{\mathbf{W},\mathbf{b}} L(\mathbf{v}, \mathbf{v}') = \sum_{i=1}^{4} \frac{1}{N_i} \sum_{j=1}^{N_i} \|\mathbf{v}_{i,j} - \mathbf{v}'_i\|^2 + \boldsymbol{\epsilon}_1 \|\mathbf{W}\|^2 \quad (2)$$

where $\boldsymbol{\epsilon}_1$ is the ℓ_2 weight decay coefficient for all the layers. N_i is the number of target faces in the age group i, and $\mathbf{v}_{i,j}$ is the j-th target face in the age group i. Unfortunately, some images in one or more positions of the aging patterns are missing, thus we cannot minimize (2) directly. However, the missing images in the aging patterns follow some statistical principles. For example, the missing image in the position "child" of the aging pattern should reflect the common traits of children. The children's skin is more smooth than that of the old people. It is very rare for children to have wrinkles on faces like the old people. Based on these statistical principles, the loss function can be further defined as follows:

$$\min_{\mathbf{W},\mathbf{b}} L(\mathbf{v}, \mathbf{v}') = \sum_{i=1}^{4} \Phi(\mathbf{v}_i, \mathbf{v}'_i) + \boldsymbol{\epsilon}_1 \|\mathbf{W}\|^2 \quad (3)$$

where $\Phi(\mathbf{v}_i, \mathbf{v}'_i)$ equals

$$\begin{cases} \frac{1}{N_i} \sum_{j=1}^{N_i} \|\mathbf{v}_{i,j} - \mathbf{v}'_i\|^2, & \mathbf{v}_i \neq \varnothing \\ \sigma^2 \mathbf{v}'^T_i \mathbf{P}_i \boldsymbol{\Lambda}^{-1} \mathbf{P}_i^T \mathbf{v}'_i + \frac{\sigma^2}{\sigma_1^2} \left\| \mathbf{v}'_i - \mathbf{P}_i \mathbf{P}_i^T \mathbf{v}'_i \right\|^2, & \mathbf{v}_i = \varnothing. \end{cases} \quad (4)$$

When the target image $\mathbf{v}_i$ at position i of the aging pattern is not empty, we still use the square error loss function. When the target image $\mathbf{v}_i$ at position i is empty, we enforce the reconstructed $\mathbf{v}'_i$ to keep the common characteristics at age group i. $\mathbf{P}_i \in \mathbb{R}^{d \times m}$ is the projection matrix calculated from the training data of group i by PCA [20]. m is the dimension of the projected subspace after PCA. m is set to 1 000 to keep $\sim$95% energy. We assume that the training images have already been normalized to zero mean and unit variance. $\boldsymbol{\Lambda} = diag[\lambda_1; \lambda_2; \ldots; \lambda_m]$ where $\lambda_1, \lambda_2, \ldots, \lambda_m$ are the m largest eigen values. $\Phi(\mathbf{v}_i, \mathbf{v}'_i)$ for $\mathbf{v}_i = \varnothing$ is inferred from probabilistic principal component analysis (PPCA) [34] and minimizing the weighted sum of these two terms is equivalent to maximizing the probability of the synthesized face in the corresponding age group.

D. Aging Face Verification Module

Face verification aims to distinguish whether a face pair has the same identity [19]. After the aging pattern synthesis module, we can obtain the synthesized faces of the input face in all the age groups (child, young, adult and old age). For each pair of reconstructed faces of a certain age group, we train a parallel CNN, which takes a face pair as the input, for the verification task. Because there are totally four age groups, we use the four synthesized face pairs and one original face pair to train five CNNs. The final verification score is obtained by fusing all the

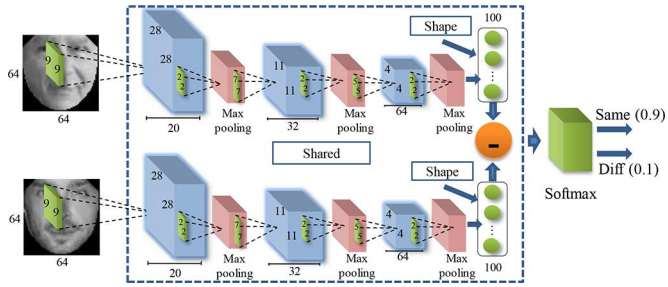


Fig. 7. Architecture of the parallel CNN. It takes a face pair as input, and predicts whether this pair belonging to the same person.

CNNs. There are two main reasons for that the original face pair is also used to train CNN. Firstly, the CNN trained from the original face pairs can still learn some strong discriminative aging-invariant features. Secondly, the synthesized faces may contain some reconstruction errors inevitably, which may be imperfect for the verification purpose. However, if the CNNs trained on reconstructed face pairs combine with that trained on the original face pairs, the complementarity among these CNNs will result in enhanced performance. Because of the synthesized faces of all the age groups, the problem of calculating the similarity of I_a and I_b has been changed to computing the similarity of $\hat{I}_a$ and $\hat{I}_b$ when $age(\hat{I}_a) = age(\hat{I}_b)$.

As shown in Fig. 7, the parallel CNN we use takes an image pair $(\hat{I}_a, \hat{I}_b)$ as the input and outputs the similarity score $s(\hat{I}_a, \hat{I}_b)$ of this pair

$$s(\hat{I}_a, \hat{I}_b) = softmax(\mathbf{W}_s |o(\hat{I}_a) - o(\hat{I}_b)| + \mathbf{b}_s) \quad (5)$$

where $o(\cdot)$ is the output of the fully-connected layer and $|\cdot|$ is element-wise absolute value. $\mathbf{W}_s$ and $\mathbf{b}_s$ are the learnable parameters in the softmax layer. $s(\hat{I}_a, \hat{I}_b)$ is the estimated probability of $(\hat{I}_a, \hat{I}_b)$ belonging the same person or different person. The goal is that $o(\cdot)$ can keep the best discriminative ability by mapping the input data into a semantic space. So, in that space, the similarity score of image pairs from the same person should be large and that from different persons should be small.

The parallel CNN structure has nine layers and is trained by stochastic gradient descent [21]. The input layer takes a pair of images as the input. The next three convolutional layers are followed by max-pooling layers to extract discriminative image features hierarchically. A nonlinear activation function is followed after the convolution operation is conducted on the input data. Here we use the rectified linear function (ReLu) [26] as the activation function. Compared with *tanh* and *sigmoid* units, ReLu converges faster and does not easily suffer from the gradient vanishing problem [21]. The followed max-pooling layer has pooling stride of 2 pixels. The following fully connected layer is learned as a semantic space, where the similarity score of image pairs from the same person is enlarged, while that from different persons is reduced. Besides the convolutional features extracted from the input face texture, the normalized coordinates of the 68 landmarks are combined as a 132 dimension vector, which is also incorporated to learn the discriminative space. The last layer is a softmax layer to produce the similarity score of the input image pair. The parameters from the input layer to the fully connection layer are shared between the input image pair.

E. Training the Whole Framework

The training process of the whole framework is summarized in Algorithm 1 and described as follows.

Algorithm 1 Training the whole framework.

Inputs: The face images and image pairs in the Training set.
Outputs: The learned parameters of the a^2-DAEs and the CNNs.
Train the a^2-DAEs:
Pretrain the encoding layers of a^2-DAE with M_0^b.
for $i = 1 \rightarrow 4$ **do**
  Pretrain the decoding layers on the i-th branch of a^2-DAE with M_i^b.
end for
Train the a^2-DAE model M_0^a with the whole training set.
for $t = 1 \rightarrow T$ **do**
  Train the t-th a^2-DAE model M_t^a.
  Obtain the synthesized aging patterns on the t-th training subset.
end for
Obtain the synthesized aging patterns of all the training subsets.
Train the parallel CNNs:
Train the CNN M_0^s on the original images.
for $i = 1 \rightarrow 4$ **do**
  Train the i-th CNN M_i^s on the constructed faces of the i-th age group.
end for
Jointly fine-tune the CNNs $\{M_i^s | i = 0 \dots 4\}$ to obtain the CNN M^s.

We employ deep belief networks (DBN) [16] to pretrain the parameters in the a^2-DAEs. DBN is stacked by RBMs as each layer and trained layerwisely [16]. The ordinary RBM models only use binary units for the visible layers, so they are not suitable for the natural images which have real pixel values. To handle this problem, the binary visible units are replaced by linear units with Gaussian noise. We totally train five DBN models $\{M_i^b | i = 0 \dots 4\}$ and use the trained parameters as the initialization of the a^2-DAEs. M_0^b is trained using all the training data, and the trained parameters are used as the initialization of the encoding layers of a^2-DAE. The a^2-DAE has four branches in the decoding layers, each of which reconstructs the faces in the specific age group. So we use $\{M_i^b | i = 1 \dots 4\}$ to pretrain the i-th branch in the decoding layers of the a^2-DAE from the training data in the i-th age group. The trained parameters in the decoding layers of $\{M_i^b | i = 1 \dots 4\}$ are used as the initialization in the decoding layers of a^2-DAE. In all the DBN models, we use *tied weights*, which means the weights in the decoding layers are the transpose of the corresponding weights in the encoding layers.

After pretraining the a^2-DAE to obtain a good initialization, we fine-tune it with the training images and their corresponding aging pattens. The detailed network structure and training process are described in Section IV-C2. To make the later parallel CNNs based on the reconstructed faces correctly trained, the overfitting of a^2-DAEs on the training set should be handled. Because the values of the loss function of a^2-DAEs on the training set can be low and those of the testing set are generally higher, the training images can "perfectly" reconstruct the aging patterns but the reconstructed aging patterns in the testing set will have relatively large reconstruction errors. This mis-match of training and testing images will result in the consequence that the later trained CNNs based on the reconstructed faces are poorly trained. To handle this problem, we train the a^2-DAEs in a cross validation way. We first train the a^2-DAE M_0^a with all the training images. Then we divide the training set into T non-overlap subsets. T is set to 6 in our experiments. M_0^a is fine-tuned with data from the training set but the i-th subset, to obtain the $\{M_i^a | i = 1 \dots T\}$ a^2-DAE. Then the aging patterns of the i-th subset are constructed from model $\{M_i^a | i = 1 \dots T\}$. In this way, we can obtain synthesized aging patterns of all the training images, during which process overfitting is well controlled. The testing synthesized faces can be constructed from the a^2-DAE M_0^a.

We use the original faces and the reconstructed aging patterns to train the parallel CNNs and fine-tune them jointly. Here "original" means real face samples from CAFE dataset, which are not synthesized from our proposed method. The reason of using the original face is explained in Section IV-D. Based on the image pairs of the original faces in the training set, we first train the CNN M_0^s. Then we use the reconstructed faces from the training pairs at the i-th age group to train the CNN $\{M_i^s | i = 1 \dots 4\}$. Then these five CNNs $\{M_i^s | i = 0 \dots 4\}$ are combined before the softmax layer. The softmax layer from each CNN is removed, and then a new softmax layer is added to combine all these CNN models. After that, the combined CNNs are fine-tuned to obtain the CNN M_f^s, which will give the final verification score.

V. Experiments

We evaluate the performance of our proposed framework and other baselines for aging face verification on our CAFE dataset. We first visualize some learned intermediate parameters. Then, we show the synthesized faces from the a^2-DAE. After that, we report the performance of face verification by quantitative evaluations. We randomly select 600 celebrities, and generate 16000 pairs (8000 pairs with same ID and 8000 pairs with different ID), among which 14000 pairs are randomly sampled out as training set and the remaining 2000 pairs are adopted as validation set. Best model parameters are selected with best validation performance. For testing, we generate 2 000 pairs (1 000 same and 1 000 different pairs) from 301 celebrities. The subjects in the training set and the testing set are mutually exclusive. We repeat this procedure of splitting training/testing sets for 5 times, and report each performance. We adopt the verification accuracy and the corresponding ROC curves for the evaluation of the proposed framework. All the deep learning based experiments are conducted on a server of 8 CPU cores and 32 GB physical memory. It is equipped with a GTX TITAN GPU of 2 688 CUDA cores and 6 GB GPU memory. The deep learning library we use is Pylearn2 [13], which is a python-based machine learning library and built on Theano [5].

In the aging pattern synthesis module, ϵ_1 is set to 0.0001. The initial learning rate is set to 0.1 and the initial momentum is 0.5. Batch size is set to 100. Around 2.5 hours are used to train each a^2-DAE of 500 epochs.

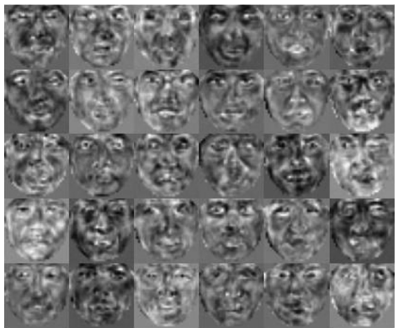


Fig. 8. Learned filters in the first encoding layers of a^2-DAE. Some parts in the filters are emphasized to capture discriminative information on the input faces.

Fig. 9. Learned filters in the last decoding layer of a^2-DAE.

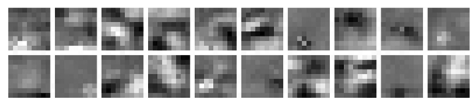


Fig. 10. Learned convolutional kernels in the first layer of CNN based on the original faces.

In the aging Face verification module, the batch size is set to 200. The initial learning rate is set to 0.01 and the initial momentum is 0.5. The number of kernels for each convolution layers are 20, 32 and 64. The kernel sizes are 9×9, 7×7 and 5×5. The fully connection layer has 100 neurons. The weight decay in this layer is set to 0.001 to balance loss values. Parameters are randomly initialized, following a gaussian distribution of mean as zero and standard deviation as 0.05. Dropout [17] probability is set to 0.2 in the fully-connected layer to avoid overfitting, which means 20% neurons will be randomly dropped during training. It will prevent neurons adapting to the training set too much. It takes around 4 hours to train each parallel CNN of 100 epochs.

At the testing stage, each face image within a testing pair is firstly passed into the aging pattern synthesis module to reconstruct aging patterns. Then, the original face pair and reconstructed face pairs of four age groups are passed through five parallel CNNs, the outputs of which are then fused to obtain the final verification score. It takes $\sim$0.5 s to process one testing face pair.

A. Visualize the Learned Parameters

We visualize the learned parameter W_1 in the first layer of a^2-DAE. Each hidden unit is fully connected with the visible units, and the parameter between them can be plotted as a face-like filter. Since there are totally 2 500 hidden units in the first layer, it is hard to plot all these parameters. We randomly select 30 of them and plot in Fig. 8. It can be seen that the plotted learned parameters are "ghost"-like faces. Some parts in the filters are emphasized, such as the corner of eyes, where the filtered faces will have larger responses. Unlike Eigen-faces [3] learned from PCA, the filters learned by a^2-DAE can perform more complex nonlinear transformations. We also plot the learned parameters of four age groups, $W'_{11}, W'_{12}, W'_{13}, W'_{14}$, in the last decoding layer. We only show the first 10 visualized faces for each branch. As shown in Fig. 9, the coefficients in the laster decoding layer look similar to the coefficients in the first encoding layer. The visualized coefficients W'_{11} look more like child faces, while W'_{14} look similar to old-age faces. This demonstrates that age difference has been captured in the a^2-DAE model.

We plot the learned kernels in the first layer of parallel CNN based on the original faces in Fig. 10. These kernels are convoluted with the input images to extract discriminative features. It can be seen that most kernels are simply edges and spots, which can effectively extract discriminative information on the human faces.

B. Synthesis Results From a^2-DAE

We visualize the reconstructed aging patterns of the testing images in Fig. 11. We plot six examples of aging patterns in three rows and two columns. In each example, the first image is the input testing image and the remaining four images are the synthesized aging patterns containing four age groups: child, young, adult and old age. The synthesized aging patterns are predicted from the trained a^2-DAE as mentioned in the previous sections. It can be seen that the faces from all the four age groups are well synthesized. The synthesized faces look very similar to the original faces, but in different age groups. The skin of the faces in the "child" group is very smooth and has no wrinkles. From "child" to "old age", the faces become less smooth, and have more wrinkles and freckles. In the "adult" and "old age" groups, some faces even have beard and mustache.

C. Quantitative Evaluation

In this subsection, we evaluate the performance of face verification. Given a pair of images, the goal is to verify whether the two images belong to the same person or not. The performance is reported as the verification accuracy and plotted on ROC curves. We compare our method with current state-of-the-art features and classifiers for general face verification, such as high dimensional local binary feature (HDLBP) [9] with probabilistic linear discriminant analysis (PLDA) [23] and probabilistic elastic matching (PEM) [22]. We also compare our method with the aging pattern subspace (AGES) [12] method, which is designed specifically for cross-age face verification. We take AGES as our cross-age baseline method, because only AGES has the ability to synthesize the faces at other age groups in the testing set, which is most similar with ours. Though the works of Suo *et al.* [31], [32] can also synthesize the aging faces, their focus is only aging synthesis, and aging face verification is not considered. So, their target is different from ours. The performances of our method and the baselines are evaluated on our CAFE dataset. The photos from each subject in the MORPH dataset have too small age gaps, which are not suitable to train

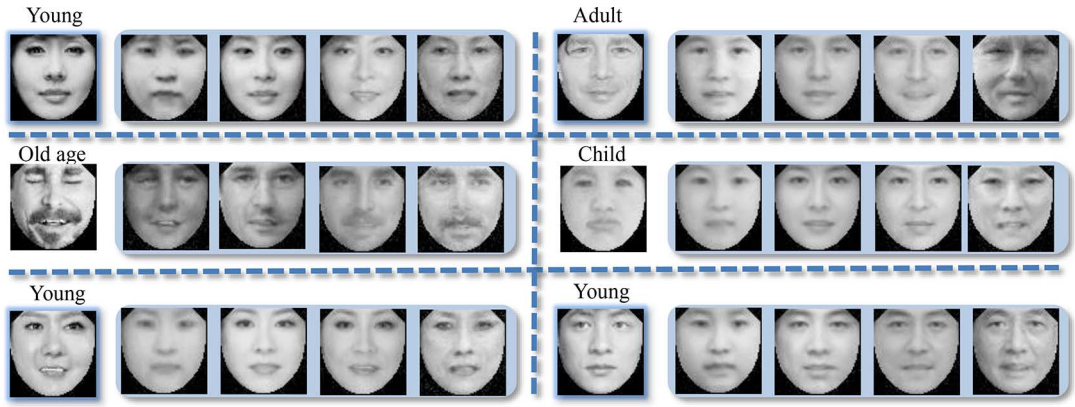


Fig. 11. Synthesized aging patterns. The first face of each group is the input face, and other four faces are synthesized faces in four age ranges. The age labels of the input faces are labeled above them.

the a^2-DAE model. The FG-NET dataset has too few subjects and images, which is not applicable for our DAFV architecture. Like other deep learning based methods, the DAFV architecture requires more samples in the training process to learn discriminative features and robust classifiers.

HDLBP: High dimensional local binary pattern (HDLBP) [9] is used as the feature and followed by probabilistic linear discriminant analysis (PLDA) [23]. Similar to the way of extracting HDLBP features in [9], we extract image patches at 27 main landmark points at five scales of image sizes: 192, 135, 96, 67, 48 from the original images of size 160×160. The patch size is set to 40×40. Each patch is divided into 4×4 cells and LBP [1] histogram is calculated in each cell. The total dimension of features is 125, 280. PCA [20] is used to reduce the feature dimension to 600, and then PLDA is used to learn a 100 dimension latent identity space, where the similarity metric of features is calculated.

PEM: In the probabilistic elastic matching (PEM) [22] method, we first crop out the center region of the image at the size 96×96. SIFT features are extracted over 3-scale image pyramid with scaling factor 0.9, from sliding window of 8×8 with 4-pixel spacing. UBM-GMM of 1 024 mixture Gaussian clusters is trained.

AGES: In this method, aging pattern subspace (AGES) [12] is learned. The testing image is projected into the aging pattern subspace, and reprojected back to get the reconstructed faces at another age. Similar to our face preprocessing method, the face images are mapped into the mean face to separate shape and texture. The face size is set to 64×64.

1) Comparison of DAFV and Other Baselines: The comparison results of our method DAFV and other baselines are shown in Table II and Fig. 12. In Fig. 12, we plot the receiver operating characteristic (ROC) curves. We fuse testing scores of all the folds, and tune the threshold from lowest score to highest. The horizontal axis is false positive rate, and vertical axis is true positive rate.

From Table II and Fig. 12, it can be seen that DAFV reaches the highest performance. For example, in Fold 1, our method reaches the accuracy of 0.7895, which is 3.55% higher than HDLBP. High dimensional LBP is one of the best hand-craft features, which has a strong discriminative capacity and shows

TABLE II
VERIFICATION ACCURACIES OF DAFV AND THE BASELINES. "AVG" MEANS AVERAGE ACCURACY ACROSS ALL FIVE FOLDS

Method	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Avg
HDLBP	0.7540	0.7640	0.7410	0.7475	0.7405	0.7494
PEM	0.7105	0.7090	0.7165	0.6950	0.7215	0.7105
AGES	0.5735	0.5670	0.5400	0.5610	0.5480	0.5579
DAFV	0.7895	0.7790	0.7530	0.7680	0.7800	0.7739

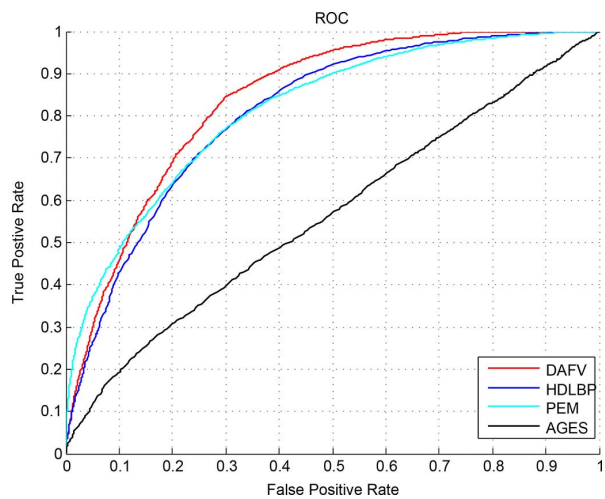


Fig. 12. ROC curve of the comparison of our method DAFV and other baselines. It can be seen that our method reaches the highest performance comparing with all the baselines.

the state-of-the-art performance in Labeled Face in the Wild (LFW) benchmark [19]. Compared with DAFV, HDLBP does not have the ability to model the aging process, which is one of the main targets in our work. Besides, HDLBP relies on the accuracy of face alignment heavily. If there are occlusions or large poses, the performance of HDLBP will drop significantly. What is more, high feature dimension of HDLBP brings huge computation cost for the training of classifier at the later stage, especially in the cases with large data scale. Though it can be reduced to lower dimension subspace to ease computation cost, the benefits from high dimensionality will also be comprimised. In contrast, DAFV propose a novel method to utilize deep learning to model the aging of faces, and produce clear visualization of the

TABLE III
ACCURACIES OF CNNS BASED ON THE SYNTHESIZED FACES IN EACH AGE GROUP OF THE FIVE FOLDS, WHICH ARE TRAINED BY THE PROPOSED CROSS-VALIDATION WAY

Group	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Avg
"child"	0.7105	0.7090	0.6895	0.7240	0.7165	0.7099
"young"	0.7295	0.7255	0.7005	0.7451	0.7075	0.7216
"adult"	0.7415	0.7260	0.7180	0.7280	0.7270	0.7281
"old-age"	0.7195	0.7105	0.6980	0.7180	0.7190	0.7130
original	0.7665	0.7595	0.7460	0.7570	0.7630	0.7584

TABLE IV
ACCURACIES OF CNNS BASED ON THE SYNTHESIZED FACES IN EACH AGE GROUP OF THE FIVE FOLDS, WHICH ARE TRAINED IN THE TRADITIONAL WAY

Group	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Avg
"child"	0.6435	0.6250	0.6345	0.6495	0.5820	0.6269
"young"	0.5650	0.5560	0.5805	0.5945	0.5820	0.5756
"adult"	0.6225	0.6060	0.6220	0.6335	0.6260	0.6220
"old-age"	0.6475	0.6535	0.6690	0.6690	0.6400	0.6558
original	0.7665	0.7595	0.7460	0.7570	0.7630	0.7584

aging effects among different aging groups given an arbitrary input face. It helps us better understand the essentials of face aging, and inspire us for better algorithms to solve this problem. Moreover, due to the good scalability of deep learning, DAFV is also suitable in cases with data of huge volume, which will improves the performance in most cases. With more identities and faces for each age group in training, better synthesis and verification results can be easily obtained.m lowest score to highest. The horizontal axis is false positive rate, and vertical axis is true positive rate.

The probabilistic elastic matching (PEM) gives worse performance than our method and HDLBP. Though PEM represents each face image as a bag of spatial-appearance features, which is robust to mis-alignment, the lack of strong prior information about the precise landmark position makes it hard to achieve the good performance as our method and HDLBP.

The AGES method has the worst performance among all the methods. This is because AGES only uses the pixel intensity as feature representation, which has a weak discriminative capacity. The similarity between the image pair is computed by the Mahalanobis distance, and no strong supervised classifiers are used for classification. Compared with it, CNN can not only extract strong discriminative features by the convolution operation in the convolutional layers for each input image, but also jointly optimize feature extraction and classification to achieve the optimal performance.

2) The Performance of CNNs Based on the Synthesized Faces: We show the performance of each CNN based on the synthesized faces of all the five folds in Tables III and IV. The a^2-DAE used in Table III is trained with our proposed cross-validation strategy, while that in Table IV is trained in a traditional way (without error control). Benefited from our proposed cross-validation training strategy, the accuracy of each single CNN based on the synthesized faces in Table III keeps relatively high performance. Averagely, CNNs in age groups "young", "adult" and "old age" even have a little higher performance than PEM. It tells that the synthesized faces maintain most of the information related to identity. The performances in "young" and "adult" are a little higher than those in "child" and "old-age", which is because that more faces in the aging patterns from "child" and "old-age" are lost than those in "young" and "adult". In contrast, the CNNs trained in the traditional way in Table IV shows much lower performance than that in Table III. With the traditional way of training, a^2-DAE suffers from heavy overfitting in training, which produces unbalanced training and testing errors. Our proposed training strategy takes the reconstruction errors into consideration, which can better control overfitting in training and testing and produces error-aware outputs.

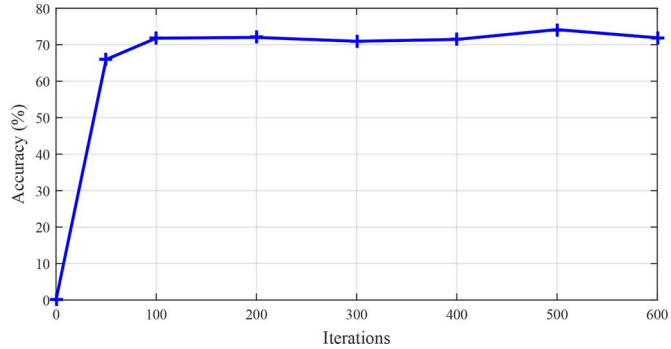


Fig. 13. Verification performance respect to different iterations in a^2-DAE model.

The quality of synthesized faces has a large influence on the verification performance at the later stage. If a^2-DAE does not completely converge, the synthesized faces would be quite noisy and look similar to the mean faces of the corresponding age group. On the other hand, if a^2-DAE is well-overfitted to the training data, it will also lead to a drop of verification performance. We plot verification performance based on the faces synthesized from a^2-DAE of different training iterations in Fig. 13. This a^2-DAE is trained from "adult" group in Fold 1, with the cross validation training strategy mentioned in the previous sections. Then all the training data are fed into a^2-DAE to get the synthesized training images in "adult" group. A parallel CNN is trained from these synthesized images, and performance is reported on 2000 validation pairs randomly sampled from the training data. In Fig. 13, the horizontal axis is the iterations of the a^2-DAE model, while the vertical axis is the verification accuracy. It can be seen that a^2-DAE quickly converge within 100 iterations. At iter-50, the verification performance is 65.95%. It shows that the verification performance is far below the current best result when a^2-DAE has not converged. From iter-100 to iter-600, the verification performance vibrates a little, and the best performance is reached at iter-500. With proper regularization and learning rate, no obvious overfitting is observed from iter-100 to iter-600, so verification performance does not drop largely.

3) Comparison of CNNs With Different Shape and Texture Combinations: The comparison of CNNs with different shape and texture combinations is shown in Table V. The performance is evaluated based on the different preprocessing of the original images. "shp/tex" means the shape and the texture are separated in the preprocessing step, but both are input into CNNs for joint optimization. The "texture" means only texture information is used as the input. "3 points" means the faces are aligned by the similarity transform from the two eye centers and the

TABLE V
ACCURACIES OF CNNS BASED ON DIFFERENT SHAPE AND TEXTURE COMBINATIONS, TRAINED FROM THE ORIGINAL FACES. "SHP/TEX" INDICATES SHAPE AND TEXTURE

Method	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Avg
shp/tex	0.7665	0.7595	0.7460	0.7570	0.7630	0.7584
texture	0.7715	0.7520	0.7485	0.7510	0.7605	0.7567
3 points	0.7455	0.7095	0.7065	0.7340	0.7365	0.7264

mouth center. It shows that "shp/tex' has a quite limited performance improvement than "texture". Restricted by the pose and expression changes in the real face images, the shape information cannot provide much discriminative information. "shp/tex' and "texture" have much higher performance than "3 points" alignment. Simply based on three landmarks of the eye centers and the mouth center, the image pixels on the faces cannot be well aligned, which will lower down the performance.

VI. CONCLUSION AND FUTURE WORK

In this work, we have developed a novel framework DAFV for aging face verification with large gaps. Two modules, aging pattern synthesis and aging face verification, are included in this framework. In the aging pattern synthesis module, we have proposed a novel deep aging-aware denoising auto-encoder (a^2-DAE) to synthesize the faces of four age groups for the input face of an arbitrary age. In the aging face verification module, given a face pair as the input, each pair of synthesized faces of the same age group is fed into a parallel CNN, and multiple parallel CNNs are fused to give the final verification score. To avoid overfitting in the aging pattern synthesis module, the cross-validation strategy is used to produce error-aware outputs. Extensive experiments on the CAFE dataset have verified the effectiveness of our proposed framework.

Our current system is definitely not perfect. We plan to further improve CAFE dataset and DAFV method in the following perspectives. Firstly, we plan to enrich CAFE dataset with more identities and face images. The synthesis and verification performance can easily benefit from the enriched data. Secondly, CAFE dataset currently only designed to study age variation, and other factors are not taken into consideration when constructing this dataset. For example, the images in CAFE dataset mostly contain near frontal faces. Without enough large pose training faces, DAFV cannot achieve good generalization ability on large pose testing set. We plan to add more faces with different poses into CAFE dataset. Pose labels will also be provided. We may change DAFV model to simultaneously synthesize aging faces and predict pose degrees to jointly model age and pose variations. Last but not least, we will also utilize face decomposition methods [15] to decompose faces into layered representation, such as face structure, face detail and face color. Then aging effects will be modeled within each layer for better synthesis results.

REFERENCES

[1] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 2037–2041, Dec. 2006.

[2] A. Albert, K. Ricanek, and E. Patterson, "The aging adult skull and face: A review of the literature and report on factors and processes of change," Univ. of North Carolina at Wilmington, Wilmington, NC, USA, Tech. Rep. WRG FSC-A, 2004.

[3] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, Jul. 1997.

[4] Y. Bengio, "Learning deep architectures for AI," *Found. Trends Mach. Learn.*, vol. 2, no. 1, pp. 1–127, 2009.

[5] J. Bergstra *et al.*, "Theano: A CPU and GPU math expression compiler," in *Proc. Python Sci. Comput. Conf.*, 2010, vol. 4.

[6] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3D faces," in *Proc. 26th Annu. Conf. Comput. Graph. Interactive Techniques*, 1999, pp. 187–194.

[7] B.-C. Chen, C.-S. Chen, and W. Hsu, "Face recognition and retrieval using cross-age reference coding with cross-age celebrity dataset," *IEEE Trans. Multimedia*, vol. 17, no. 6, pp. 804–815, Jun. 2015.

[8] B.-C. Chen, Y.-Y. Chen, Y.-H. Kuo, and W. H. Hsu, "Scalable face image retrieval using attribute-enhanced sparse codewords," *IEEE Trans. Multimedia*, vol. 15, no. 5, pp. 1163–1173, Aug. 2013.

[9] D. Chen, X. Cao, F. Wen, and J. Sun, "Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2013, pp. 3025–3032.

[10] J. Y. Choi, W. De Neve, K. N. Plataniotis, and Y. M. Ro, "Collaborative face recognition for improved face annotation in personal photo collections shared on online social networks," *IEEE Trans. Multimedia*, vol. 13, no. 1, pp. 14–28, Feb. 2011.

[11] M. de Berg, O. Cheong, M. van Kreveld, and M. Overmars", *Computational Geometry: Algorithms and Applications*, 3rd ed. Berlin, Germany: Springer-Verlag, 2008.

[12] X. Geng, Z.-H. Zhou, and K. Smith-Miles, "Automatic age estimation based on facial aging patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 12, pp. 2234–2240, Dec. 2007.

[13] I. J. Goodfellow *et al.*, "Pylearn2: A machine learning research library," *CoRR*, 2013 [Online]. Available: http://arxiv.org/abs/1308.4214

[14] A. Goshtasby, "Piecewise linear mapping functions for image registration," *Pattern Recog.*, vol. 19, no. 6, pp. 459–466, 1986.

[15] D. Guo and T. Sim, "Digital face makeup by example," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2009, pp. 73–79.

[16] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[17] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *CoRR*, 2012. [Online]. Available: http://arxiv.org/abs/1207.0580

[18] C.-K. Hsieh, S.-H. Lai, and Y.-C. Chen, "Expression-invariant face recognition with constrained optical flow warping," *IEEE Trans. Multimedia*, vol. 11, no. 4, pp. 600–610, Jun. 2009.

[19] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Univ. of Massachusetts, Amherst, MA, USA, Tech. Rep. 07-49, 2007.

[20] I. Jolliffe, "Principal component analysis," in *Encyclopedia of Statistics in Behavioral Science*. New York, NY, USA: Wiley, 2002.

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Adv. Neural Inf. Process. Syst.*, pp. 1097–1105, 2012.

[22] H. Li, G. Hua, Z. Lin, J. Brandt, and J. Yang, "Probabilistic elastic matching for pose variant face verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2013, pp. 3499–3506.

[23] P. Li, Y. Fu, U. Mohammed, J. H. Elder, and S. J. Prince, "Probabilistic models for inference about identity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 1, pp. 144–157, Jan. 2012.

[24] Z. Li, U. Park, and A. K. Jain, "A discriminative model for age invariant face recognition," *IEEE Trans. Inf. Forensics Security*, vol. 6, no. 3, pp. 1028–1037, Sep. 2011.

[25] P. Luo, X. Wang, and X. Tang, "Hierarchical face parsing via deep learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 2480–2487.

[26] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. ICML*, 2013, vol. 30.

[27] U. Park, Y. Tong, and A. K. Jain, "Age-invariant face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 5, pp. 947–954, May 2010.

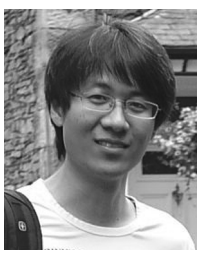
[28] K. Ricanek, Jr. and T. Tesafaye, "Morph: A longitudinal image database of normal adult age-progression," in *Proc. 7th Int. Conf. Automat. Face Gesture Recog.*, 2006, pp. 341–345.

[29] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," *Adv. Neural Inf. Process. Syst.*, pp. 1988–1996, 2014.
[30] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10 000 classes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2014, pp. 1891–1898.
[31] J. Suo, X. Chen, S. Shan, W. Gao, and Q. Dai, "A concatenational graph evolution aging model," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2083–2096, Nov. 2012.
[32] J. Suo, S.-C. Zhu, S. Shan, and X. Chen, "A compositional and dynamic model for face aging," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 3, pp. 385–401, Mar. 2010.
[33] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2014, pp. 1701–1708.
[34] X. Tan and B. Triggs, "Enhanced local texture feature sets for face recognition under difficult lighting conditions," *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1635–1650, Jun. 2010.
[35] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," *J. Royal Statist. Soc.*, ser. B (Statist. Methodol.), vol. 61, no. 3, pp. 611–622, 1999.
[36] H. Wang, S. Z. Li, and Y. Wang, "Face recognition under varying lighting conditions using self quotient image," in *Proc. 6th IEEE Inte. Conf. Automat. Face Gesture Recog.*, May 2004, pp. 819–824.
[37] T. Wu and R. Chellappa, "Age invariant face verification with relative craniofacial growth model," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 58–71.
[38] Z. Zhu, P. Luo, X. Wang, and X. Tang, "Deep learning multi-view representation for face recognition," *CoRR*, 2014 [Online]. Available: http://arxiv.org/abs/1406.6947

Luoqi Liu is currently working toward the Ph.D. degree in electrical and computer engineering at the National University of Singapore, Singapore.

His research interests include computer vision and multimedia.

Mr. Liu was the recipient of the Best Paper Award from ACM MM 2013 and the Best Student Paper Award (Gold Prize) from PREMIA'14.



Chao Xiong received the M.Sc. degree in communication and signal processing from Imperial College London, London, U.K., in 2011, and is working toward the Ph.D. degree in a joint program between Imperial College London, London, U.K., and the National University of Singapore, Singapore.

His research interests include face recognition in computer vision.

Hanwang Zhang is currently a Research Fellow with the School of Computing, National University of Singapore (NUS), Singapore. His research interest includes multimedia and computer vision, developing techniques for efficient search and recognition in visual contents.

Mr. Zhang was the recipient of the Best Demo Runner-Up Award from ACM MM 2012, the Best Student Paper Award from ACM MM 2013, and the Best Ph.D. Thesis Award from the School of Computing, NUS, in 2014.

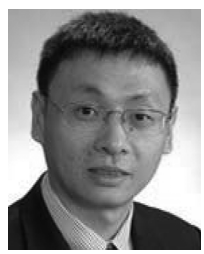
Zhiheng Niu received the B.S., M.S., and Ph.D. degrees from the Harbin Institute of Technology (HIT), Harbin, China, in 2003, 2005, and 2009, respectively.

He was a Senior R&D Engineer with the Panasonic Research and Development Center Singapore (PRDCSG), Singapore, from 2009 to 2013. He was a Senior Research Fellow with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore, from 2013 to 2015. His research interest includes object detection, object tracking, pattern classification, deep learning, and facial image analysis.

Meng Wang received the Ph.D. degree in electronic engineering and information science from the University of Science and Technology of China (USTC), Hefei, China, in 2008.

He previously worked as an Associate Researcher with Microsoft Research Asia, Beijing, China, and then as a core member in a startup in silicon valley. After that, he worked with the National University of Singapore as a Senior Research Fellow. He is currently a Professor with the Hefei University of Technology, Heifei, China. He has authored or coauthored more than 150 book chapters and journal and conference papers. His current research interests include multimedia content analysis, search, mining, recommendation, and large-scale computing.

Prof. Wang was the recipient of the Best Paper Award from both the 17th and 18th ACM International Conference on Multimedia, the Best Paper Award from the 16th International Multimedia Modeling Conference, the Best Paper Award from the 4th International Conference on Internet Multimedia Computing and Service, and the Best Demo Award from the 20th ACM International Conference on Multimedia.



Shuicheng Yan (M'06–SM'09) is currently a (Dean's Chair) Associate Professor with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore. He has authored or coauthored over 370 technical papers over a wide range of research topics, with Google Scholar citation >21 000 times and H-index-61. His research interests include computer vision, multimedia, and machine learning.

Dr. Yan was or will be General or Program Co-Chair of MMM'13, PCM'13, ACM MM'15, ICMR'17, and ACM MM'17. He was the recipient of the Best Paper Award from ACM MM'13 (Best Paper and Best Student Paper), ACM MM'12 (demo), PCM'11, ACM MM'10, ICME'10, and ICIMCS'09, the winning prize of the classification task in PASCAL VOC 2010–2012, the winning prize of the segmentation task in PASCAL VOC'12, the Honorable Mention prize of the detection task in PASCAL VOC'10, 2010 TCSVT Best Associate Editor (BAE) Award, 2010 Young Faculty Research Award, 2011 Singapore Young Scientist Award, and 2012 NUS Young Researcher Award. He was the coauthor of the papers that was the recipient of the Best Student Paper Award of PREMIA'09, PREMIA'11, PREMIA'12, PREMIA'14, and PREMIA'15.