A Maximum Entropy Feature Descriptor for Age Invariant Face Recognition

Dihong Gong[1]
dh.gong@siat.ac.cn

Zhifeng Li[1]
zhifeng.li@siat.ac.cn

Dacheng Tao[2]
dacheng.tao@uts.edu.au

Jianzhuang Liu[3,4]
liu.jianzhuang@huawei.com

Xuelong Li[5]
xuelong_li@opt.ac.cn

[1]Shenzhen Key Lab of Computer Vision and Pattern Recognition
Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China
[2]Centre for Quantum Computation & Intelligent Systems, Faculty of Engineering and IT, University of Technology, Sydney, NSW 2007, Australia
[3]Dept. of Information Engineering, the Chinese University of Hong Kong
[4]Media Lab, Huawei Technologies Co. Ltd., China
[5]Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences

Abstract

In this paper, we propose a new approach to overcome the representation and matching problems in age invariant face recognition. First, a new maximum entropy feature descriptor (MEFD) is developed that encodes the microstructure of facial images into a set of discrete codes in terms of maximum entropy. By densely sampling the encoded face image, sufficient discriminatory and expressive information can be extracted for further analysis. A new matching method is also developed, called identity factor analysis (IFA), to estimate the probability that two faces have the same underlying identity. The effectiveness of the framework is confirmed by extensive experimentation on two face aging datasets, MORPH (the largest public-domain face aging dataset) and FGNET. We also conduct experiments on the famous LFW dataset to demonstrate the excellent generalizability of our new approach.

1. Introduction

Age invariant face recognition (AIFR) is an important but challenging area of face recognition research. AIFR is useful in a number of practical applications, for example finding missing children and identifying criminals based on their mug shots [1, 2]. Although significant progress has been made in face recognition, AIFR still remains a major challenge in real world applications such as face recognition systems, in which age-related face image analysis has most traction.

Most existing works focus on age estimation [3–13] and aging simulation [14–18]; only a very limited number of studies tackle AIFR [19, 20]. A typical AIFR approach is to use face modeling to synthesize and render face images to the same age as the gallery image prior to recognition [4, 14, 18, 33]. However, due to strong parametric assumptions and the complexity of the algorithm, these methods are computationally expensive and the results are often unstable for real-world face recognition.

Recently, several discriminative methods have been proposed to improve the performance of AIFR [19, 20, 34, 35, 41, 42]. For example, the method in [19] uses gradient orientation pyramids (GOPs) for feature representation combined with SVMs to verify faces as they age. The method presented in [20] combines scale-invariant feature transforms (SIFT) [23] and multi-scale local binary patterns (MLBP) [26] with a random sampling-based fusion framework to improve AIFR performance. Random sampling linear discriminant analysis (LDA) variants have also been proposed [34][35] to tackle the face-aging problem in face recognition, which have been shown to be robust, have fewer parameter and training data requirements, and outperform existing methods. More recently, [41] and [42] have notably improved the performance of AIFR.

In this paper, we propose a new two-step AIFR approach. First, a new feature descriptor called the maximum entropy feature descriptor (MEFD) is used to extract expressive and informative features. Unlike existing feature descriptors, MEFD can maximize the expressive power in terms of maximum entropy, which is highly beneficial to classification. Second, a factor analysis-based matching framework termed identity factor analysis (IFA) is developed to further improve recognition performance. We demonstrate the efficiency of the new approach compared to state-of-the-art methods in several face databases including MORPH (the largest public-domain facial aging dataset) [30], FGNET [36], and the famous LFW database [50].

2. Proposed Approach

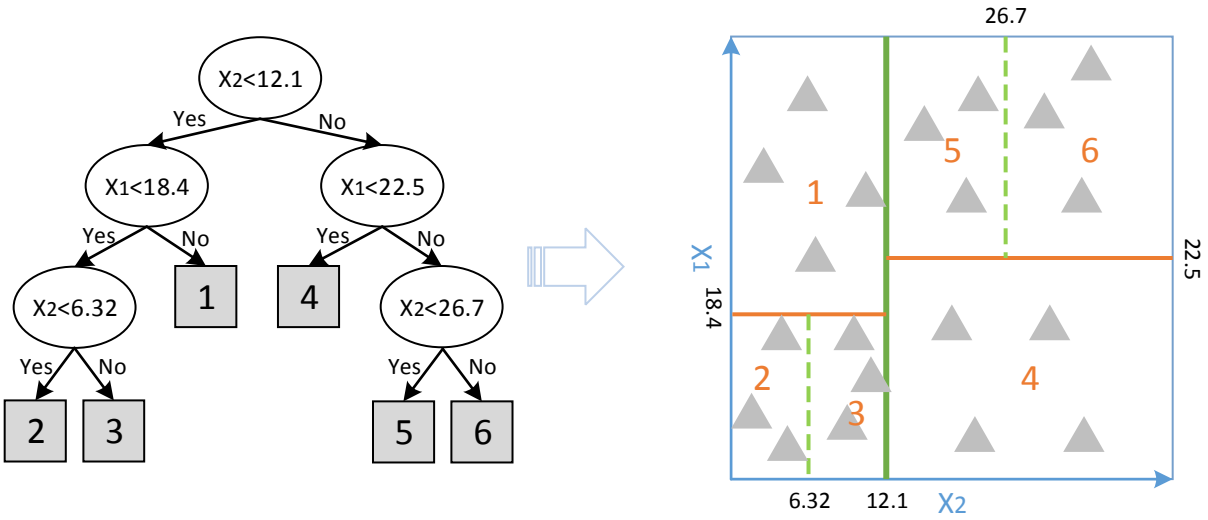

Figure 1. The decision tree based encoding scheme. The internal nodes are decision rules, while the leaf nodes represent code assignment. The decision tree is trained in a greedy manner such that each leaf node hosts a similar number of training samples, as illustrated on the right. X1 and X2 represent two different attributes.

2.1. Proposed Feature Descriptor

Of the existing local image descriptors, LBPs [21][25][26] have been shown to be very effective [22]. However, the LBP descriptor has its limitations. First, it empirically assumes that the uniform binary patterns have higher emergence frequency than the non-uniform binary patterns. However, this is not always the case. In the context of AIFR, we have found that some uniform binary patterns rarely appear, while the appearance frequency of some non-uniform binary patterns is high, as illustrated in Fig. 3. This phenomenon makes the descriptor less informative (in terms of entropy) since the uniform scheme assigns single codes to low frequency patterns while high-frequency non-uniform patterns are combined into a single code. Second, the local feature length for uniform LBP must be fixed at 59, which may limit its use.

In order to overcome these limitations, we propose a learning-based coding scheme to convert binary patterns into specific codes. Unlike many handcrafted encoders, in our approach the encoder is specifically trained using a set of training face images such that the frequency of output codes distributes as evenly as possible; this maximizes the discriminative ability in terms of maximum entropy. As illustrated in Figure 1, the pattern space is quantized using a decision tree. For instance, suppose we have a training dataset $X = \{x_i | x_i \in R^{d\times 1}, i = 1, \ldots, N\}$; the goal of the decision tree is to build a partition-based model that assigns each point x_i with a code $y_i \in \{1,2,\ldots,K\}$, where the probability mass function over the set of codes is as close to a uniform distribution as possible.

This decision tree grows in a greedy manner such that, at each split step, it extends the *best* node to maximize the entropy of the code distribution. Suppose we are extending a tree of K leaf nodes by splitting the i-th leaf node to produce a new tree of (K+1) leaf nodes whose entropy is:

$$E_i^{K+1} = -\left(\sum_{k=1}^{i-1} p(k)\log p(k) + \sum_{k=i+1}^{K} p(k)\log p(k) + p1\log p1 + p2\log p2\right), \quad (1)$$

where $p1$ and $p2$ represent the probabilities of a pattern falling into these new partitions. The E_i^{K+1} can be rewritten as:

$$E_i^{K+1} = E_i^K + p(i)\log(i) - (p1\log p1 + p2\log p2). \quad (2)$$

In order to optimize (2), we maximize the information gain:

$$G(i) = p(i)\log(i) - (p1\log p1 + p2\log p2). \quad (3)$$

For a given i, the p(i) is fixed, and thus we can maximize the G(i) by splitting the node in such a way that the two new partitions are as even as possible. In our experiments, we found that splitting the node with the mean of attributes of samples falling into partition i (referred to as the *mean-split-point*) is satisfactory. Thus, at each node splitting step, we evaluate the optimal information gain of each leaf node by calculating the information gain obtained at the *mean-split-point* for each attribute; the node of maximum information gain is then split at the *mean-split-point* with respect to the corresponding attribute. The decision-tree based encoder is described in detail in Algorithm 1.

Algorithm 1 Learning the encoding tree

Inputs: A set of training images $I=\{t_i \mid i=1,...,N\}$, sampling radii = *r,* and number of codes *K*.

1. **Pixel vector extraction.** For each pixel in the training images, compute its pixel vector by subtracting its 8-neighbor pixels with radii *r* to form a pixel vector set $X=\{\vec{f}_k \in R^{8\times 1} \mid k=1,...,M\}$.

2. **Tree initialization.** Insert a root node hosting all training pixel vectors with *prob* = 1.0.

3. **Recursive tree extension.** For it = 2 to K:

   (i) for each leaf node in the tree, compute its information gain using the *mean-split-point*.

   (ii) extend the tree by splitting the leaf node with maximum information gain.

4. **Code assignment.** For each leaf node, assign a distinct code to it ranging from 1 to K.

Outputs: The encoding tree of K leaf nodes.

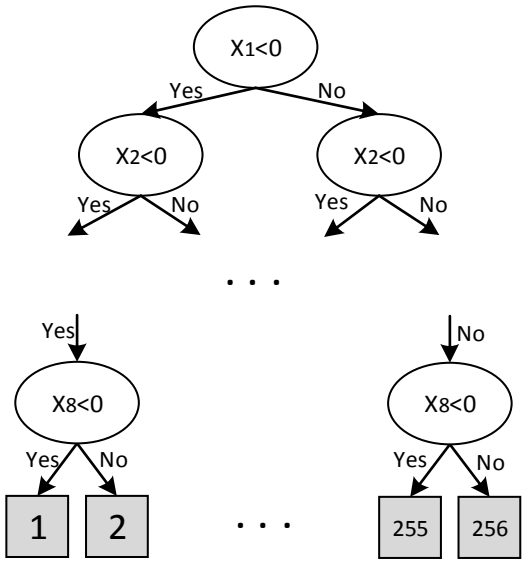

Figure 2. The LBP equivalent of the decision tree-based feature descriptor. The internal decision nodes use "0" as decision points with each level using attributes 1, 2,…, 8, corresponding to a complete binary tree.

A set of uniformly distributed codes can be obtained using this learning-based encoding scheme, and the histogram of these codes can be used as a descriptor. In this way, the new maximum entropy feature descriptor (MEFD) can be designed. Interestingly, the LBP encoder can be viewed as a special case of our new encoder: the LBP encoder corresponds to a complete decision tree of height = 8 with 256 leaf nodes, and each internal node uses "0" as the split-point, as illustrated in Fig. 2. Compared to the LBP descriptor, the code histogram of the MEFD is more uniformly distributed across different people, as illustrated in Fig. 3. This leads to high discriminative power and significantly boosts recognition performance (see Section 3). In addition, the length of the local features can be varied to improve the tradeoff between discriminative power and model complexity.

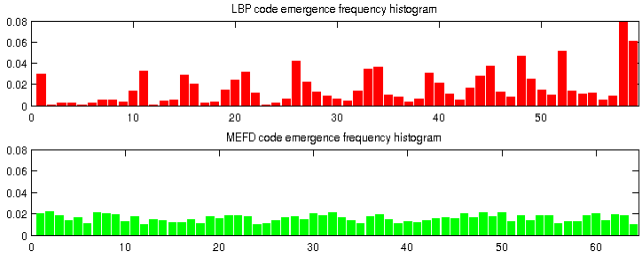

Figure 3. An illustration of the code frequency histogram for the U2 encoding scheme [21] and the proposed encoding scheme with 64 codes. The frequency is counted over 20,000 facial images from 10,000 people (with each person having two facial images with the largest age gap) in the MORPH Album 2 dataset. The histogram distribution of the MEFD is clearly much more uniform than that of LBP.

Discriminative information can be extracted from the facial image for further processing using MEFD. Since the entire face image (which has high structural complexity) is difficult to characterize using a single image descriptor, we use a patch-based local feature representation scheme (also called densely sampled local feature description) in this paper. The input face image is first divided into a set of overlapping patches, after which the new descriptor is applied to each patch to extract discriminant information. In order to ensure local consistency, a 50% overlap between adjacent patches is used. The detailed procedure is as follows:

(1) The whole face images are divided into a set of overlapping patches. A face image of size HxW is divided into a set of s overlapping patches (of size 16x16 pixels) that overlap by r pixels. The number of horizontal (M) and vertical (N) patches obtained are

$$N=(W-s)/r+1 \quad (4)$$

$$M=(H-s)/r+1 \quad (5)$$

(2) For each M×N patch, MEFD is applied to describe the microstructure of this region at multiple scales with radii {1, 3, 5, 7}.

(3) For each patch, a d-dimensional feature vector is obtained. These feature vectors are concatenated into a single MxNx4xd-dimensional feature vector for a given face image.

2.2. Dimensionality Reduction

Due to the use of densely sampling technique and multi-scale scheme in feature representation stage, the extracted MEFD feature vector is of high dimension. So we need to perform dimension reduction prior to feature matching stage. Inspired by the feature slicing technique in [51][52][53], we first divide the extracted long feature vector into several slices equally, and then apply PCA+LDA [27] on each slice to obtain a compressed slice of smaller feature vector for subsequent analysis.

2.3. Matching Framework

In this section, we develop an effective matching framework called identity factor analysis (IFA) for feature classification. It has been shown that the observable face features $\vec{t}$ (in the presence of aging variations) can be decomposed into the following four terms: the mean component ($\vec{\beta}$), the identity-related component ($U\vec{x}$), the age-related component ($V\vec{y}$), and the noise component ($\vec{\varepsilon}$). This decomposition model can be formulated as:

$$\vec{t} = \vec{\beta} + U\vec{x} + V\vec{y} + \vec{\varepsilon}. \quad (6)$$

The terms in the model are as follows:

1. $\vec{t}$ is a $d \times 1$ random vector representing observable feature vectors from the face image.
2. $\vec{\beta}$ is a $d \times 1$ vector representing the mean of the face features.
3. $\vec{x}$ is a $p \times 1$ random vector representing the identity-hidden factor with a prior distribution of $N(0, I)$.
4. $\vec{y}$ is a $q \times 1$ random vector representing the age-hidden factor with a prior distribution of $N(0, I)$.
5. $\vec{\varepsilon}$ is a $d \times 1$ random vector representing the additive noise. The noise is modeled as isotropic Gaussian where $\vec{\varepsilon} \sim N(0, \sigma^2 I)$.
6. U is a $d \times p$ matrix whose columns comprise a subspace capturing the *identity variations*.
7. V is a $d \times q$ matrix whose columns comprise a subspace capturing the *age variations*.

The model parameters $\{\vec{\beta}, U, V, \vec{\varepsilon}\}$ can be estimated using the EM algorithm. Such a decomposition model is quite effective for age invariant face recognition. We can use an experiment to illustrate this point. We randomly select 500 pairs of gallery samples and probe samples from the MORPH Album2 dataset [30], and then compute their cosine distance values in feature space (the original MEFD feature space) and the latent space (the space associated with the identity-related components) respectively, as plotted in Figure 4. It is very clear that the intra-class variations between the gallery sample and the probe sample can be better reduced in the latent space than the original feature space.

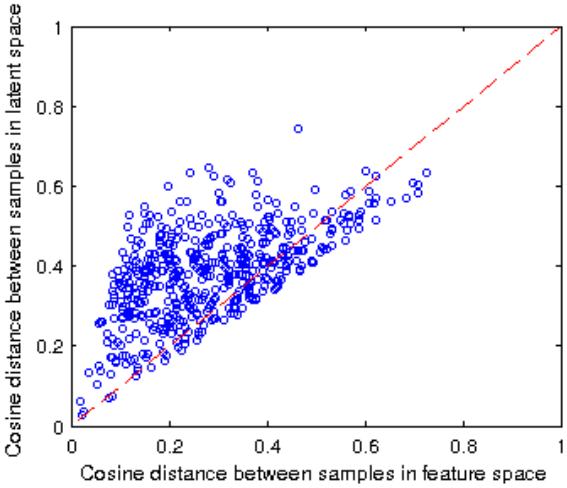

Figure 4. Illustration of the distribution of the cosine distance values of the gallery samples and the probe samples in feature space and the latent space.

Based on the decomposition model, we can perform face recognition in the presence of age variations. A typical approach is to use the identity-related component as the age invariant features and hence calculate the cosine distance of the identity components ($U\vec{x}$) of the gallery sample and the probe sample for direct classification [41]. However, note that the observed face feature is often noisy, and it is therefore very difficult to accurately estimate the identity component in real face recognition applications. In view of this, we develop a new matching method called identity factor analysis. Instead of asking *what* the identity is, our new method asks whether the gallery and probe sample are from the same identity.

Denote $\vec{t}_g$ the gallery sample and $\vec{t}_p$ the probe sample. If $\vec{t}_g$ and $\vec{t}_p$ are from the same identity, according to (6) we have

$$\begin{bmatrix} \vec{t}_g \\ \vec{t}_p \end{bmatrix} = \begin{bmatrix} \vec{\beta} \\ \vec{\beta} \end{bmatrix} + \begin{bmatrix} U & V & 0 \\ U & 0 & V \end{bmatrix} \begin{bmatrix} \vec{x}_{g,p} \\ \vec{y}_g \\ \vec{y}_p \end{bmatrix} + \begin{bmatrix} \vec{\varepsilon}_g \\ \vec{\varepsilon}_p \end{bmatrix}.$$

The above formula can also be represented as:

$$\bar{T} = \bar{\beta} + \bar{A}_1 \bar{Z}_1 + \bar{\varepsilon}_{g,p}. \quad (7)$$

The probability density function of $\bar{T}$ is

$$p(\bar{T}) = N\left(\bar{T} \mid \bar{\beta}, \bar{A}_1 \bar{A}_1^T + \bar{\Sigma}\right)$$

$$= N\left(\begin{bmatrix} \vec{t}_g \\ \vec{t}_p \end{bmatrix} \middle| \begin{bmatrix} \vec{\beta} \\ \vec{\beta} \end{bmatrix}, \begin{bmatrix} UU^T + VV^T + \sigma^2 I & UU^T \\ UU^T & UU^T + VV^T + \sigma^2 I \end{bmatrix}\right)$$

$$= N\left(\begin{bmatrix}\vec{t}_g\\ \vec{t}_p\end{bmatrix}\middle|\middle|\begin{bmatrix}\vec{\beta}\\ \vec{\beta}\end{bmatrix}, \begin{bmatrix}\Sigma_{tot} & \Sigma_{ac}\\ \Sigma_{ac} & \Sigma_{tot}\end{bmatrix}\right), \text{ where}$$

$$\overline{\Sigma} = \begin{bmatrix}\sigma^2 I & 0\\ 0 & \sigma^2 I\end{bmatrix}, \Sigma_{tot} = UU^T + VV^T + \sigma^2 I, \Sigma_{ac} = UU^T.$$

If $\vec{t}_g$ and $\vec{t}_p$ are from different identities, then we have

$$\begin{bmatrix}\vec{t}_g\\ \vec{t}_p\end{bmatrix} = \begin{bmatrix}\vec{\beta}\\ \vec{\beta}\end{bmatrix} + \begin{bmatrix}U & 0 & V & 0\\ 0 & U & 0 & V\end{bmatrix}\begin{bmatrix}\vec{x}_g\\ \vec{x}_p\\ \vec{y}_g\\ \vec{y}_p\end{bmatrix} + \begin{bmatrix}\vec{\varepsilon}_g\\ \vec{\varepsilon}_p\end{bmatrix}.$$

The above formula can also be represented as:

$$\overline{T} = \overline{\beta} + \overline{A}_2 \overline{Z}_2 + \overline{\varepsilon}_{g,p}.$$

The probability density function of $\overline{T}$ is

$$p(\overline{T}) = N\left(\overline{T} \mid \overline{\beta}, \overline{A}_2 \overline{A}_2^T + \overline{\Sigma}\right)$$

$$= N\left(\begin{bmatrix}\vec{t}_g\\ \vec{t}_p\end{bmatrix}\middle|\middle|\begin{bmatrix}\vec{\beta}\\ \vec{\beta}\end{bmatrix}, \begin{bmatrix}UU^T + VV^T + \sigma^2 I & 0\\ 0 & UU^T + VV^T + \sigma^2 I\end{bmatrix}\right)$$

$$= N\left(\begin{bmatrix}\vec{t}_g\\ \vec{t}_p\end{bmatrix}\middle|\middle|\begin{bmatrix}\vec{\beta}\\ \vec{\beta}\end{bmatrix}, \begin{bmatrix}\Sigma_{tot} & 0\\ 0 & \Sigma_{tot}\end{bmatrix}\right).$$

The matching score is then calculated as:

$$LR(\vec{t}_g, \vec{t}_p) = \log \frac{N\left(\begin{bmatrix}\vec{t}_g\\ \vec{t}_p\end{bmatrix}\middle|\middle|\begin{bmatrix}\vec{\beta}\\ \vec{\beta}\end{bmatrix}, \begin{bmatrix}\Sigma_{tot} & \Sigma_{ac}\\ \Sigma_{ac} & \Sigma_{tot}\end{bmatrix}\right)}{N\left(\begin{bmatrix}\vec{t}_g\\ \vec{t}_p\end{bmatrix}\middle|\middle|\begin{bmatrix}\vec{\beta}\\ \vec{\beta}\end{bmatrix}, \begin{bmatrix}\Sigma_{tot} & 0\\ 0 & \Sigma_{tot}\end{bmatrix}\right)}$$

$$= const + 0.5\vec{t}_g{}^T Q \vec{t}_g + 0.5\vec{t}_p{}^T Q \vec{t}_p + \vec{t}_g{}^T P \vec{t}_p, \quad (8)$$

where

$$P = \Sigma_{tot}^{-1}\Sigma_{ac}(\Sigma_{tot} - \Sigma_{ac}\Sigma_{tot}^{-1}\Sigma_{ac})^{-1}, \; Q = \Sigma_{tot}^{-1} - (\Sigma_{tot} - \Sigma_{ac}\Sigma_{tot}^{-1}\Sigma_{ac})^{-1},$$

$$const = \frac{1}{2}\log\begin{vmatrix}\Sigma_{tot} & 0\\ 0 & \Sigma_{tot}\end{vmatrix} - \frac{1}{2}\log\begin{vmatrix}\Sigma_{tot} & \Sigma_{ac}\\ \Sigma_{ac} & \Sigma_{tot}\end{vmatrix}.$$

Based on (8), the matching score of the gallery and probe samples can be estimated, resulting in robust face recognition.

The details of the entire algorithm can be summarized as follows:

(1) For any face image, first extract MEFD features using the method described in Section 2.1.

(2) The extracted MEFD features have high feature dimensions. Therefore, divide them into several equal slices and apply PCA + LDA [27] to each slice for dimension reduction.

(3) Apply the IFA model to each slice to obtain the matching score (Equation 8). Combine the matching scores of all the slices to obtain a final decision using the sum rule.

2.4. Discussion

Histogram-based feature descriptors encode visual clues into pattern distributions. For example, the SIFT descriptor encodes gradient information while the LBP descriptor encodes orientation information. The extracted features (the pattern distributions) can be viewed as random variables in a *D*-dimensional feature space, where *D* is the number of different patterns. According to information theory, entropy is a measure of the uncertainty in a random variable, and it quantifies the expected value of the information contained in a message.

Compared to the popular handcrafted feature descriptors, e.g., SIFT, LBP, and HOG, our encoding mechanism not only encodes the orientation information (by assigning a code to a pixel based on the difference between the pixel and its neighboring pixels in different orientations, similar to LBP), but simultaneously maximizes the information contained in the encoded feature space. The richness of the information contained in the feature space reflects the expressive power of the feature descriptor: the more information it contains, the more objects it can describe. As a toy example, consider the extreme case is which a descriptor has 100% occurrence probability for a specific pattern and 0% probability for all other patterns; in this case, it does not contain any information since it cannot distinguish two objects with the same pattern distributions. In this sense, our approach can learn a feature descriptor that maximizes the expressive power, resulting in a more informative and expressive descriptor. Another merit of our new feature descriptor is that it is easy to implement and can easily be combined with existing feature descriptors to further boost recognition performance.

Our approach differs significantly from [41]. In the feature representation stage, [41] uses the HOG features as a feature presentation, whereas in this paper, we develop a new feature descriptor called MEFD to better represent the face image at different ages. In the feature matching stage, [41] tries to estimate the identity-related component as the age invariant features, and hence calculate the cosine distance of the identity components of the gallery sample and the probe sample for direct classification. Unlike [41], this paper develops a new probabilistic matching framework called IFA to estimate the probability that the two faces have the same underlying identity. Thus, our matching framework is more robust, as supported by our experimental results.

Both the proposed feature descriptor and the matching framework can be extended to address the general face recognition problem. For general application, the aging term ($V\vec{y}$) in equation (6) can be replaced with general intra-personal variations (age, pose, illumination, expression, etc.), as supported by the experimental results in the next section.

3. Experiments

3.1. Experiments on the MORPH Album 2 dataset.

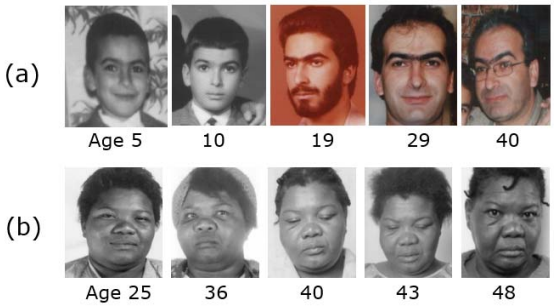

Figure 5. Example images from (a) FGNET [36] and (b) MORPH [30].

There are two well-known public-domain face aging databases: FGNET [36] and MORPH [30]. Some examples of them are shown in Figure 5. FGNET is a relatively small database consisting of 1002 face images from 82 different people. MORPH [30] contains two separate datasets: Album 1 and Album 2. MORPH Album 1 only contains 1690 face images from 625 different people, while MORPH Album 2 is much larger; we therefore conduct our experiments on an extended version of MORPH Album 2 (the largest publicly available face aging dataset) [30] for large-scale experimental validation. This dataset contains about 78,000 face images from 20,000 different people. It is partitioned into a training set and an independent test set. For training, we selected 20,000 face images from 10,000 subjects with each subject represented by two images with the largest age gap. For testing, a gallery set and a probe set were collected from the remaining 10,000 subjects: the gallery set contains 10,000 face images corresponding to the youngest age of the subjects, while the probe set contains 10,000 face images corresponding to the oldest age of the subjects. All facial images were automatically preprocessed as follows: (1) the face images were rotated to align them to vertical; (2) the face images were scaled so that the distance between the eyes was equal in all the images; and (3) the face images were cropped to 200 × 150 to remove the background and hair regions.

Parameter exploration. Here, 10-fold cross validation was used to determine the parameter K (length of the local features). 10,000 training images were randomly partitioned into 10 subsets of equal size. A single subset was retained as validation data to test the model, and the remaining 9 subsets were used as training data. The process was repeated 10 times and the average performance reported (Fig. 6). It can be seen that there is a tradeoff between a small and large K: a smaller K yields a simpler model with better generalizability but less discriminative power, while a larger K is more discriminative but less generalizable. The optimal tradeoff between accuracy and compactness was K = 64 (Fig. 6), which was used in subsequent experiments.

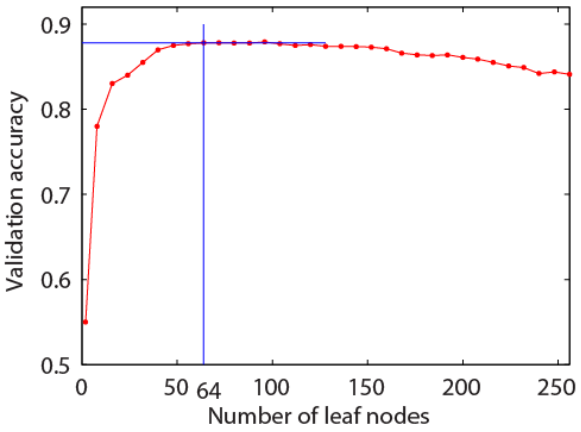

Figure 6. Validation accuracies vs. #leaf nodes.

Comparison with state-of-the-art feature descriptors. We first investigated the effectiveness of the proposed MEFD by comparing it with state-of-the-art descriptors used for face recognition. In this experiment, we used the features (extracted by these descriptors) directly for matching. The matching scheme was the cosine distance. All the other methods listed in Table 1 were implemented using the parameters suggested in their original papers. The LBP feature descriptor is the original LBP descriptor. The multi-scale LBP (MLBP) feature descriptor is an extension of LBP that computes LBP descriptors at four different radii {1, 3, 5, 7} [21]. The SIFT feature descriptor [23] quantizes both the spatial location and orientation of the image gradient within an image patch and computes a histogram in which each bin corresponds to a specific spatial location and gradient orientation location; SIFT has been widely used in face recognition [24][20]. The SIFT-Rank [40] algorithm is a revised version of SIFT that uses SIFT ranking values as features. In our experiment, the SIFT features were extracted on the same landmark points as our descriptor to produce 128-dimensional local features corresponding to 4x4 cells and 8 bins. For fair comparison, we also included the multi-scale version of SIFT with sampling scales {1,3,5,7}. Bio-inspired features (BIF) is a recently developed descriptor that has been successfully applied by the face aging community to applications including face-based human age estimation [6]. The comparative results are shown in Table 1; it can be seen that our approach outperforms the other descriptors by a clear margin.

Table 1. Comparison with popular feature descriptors.

Feature descriptors	Recognition Accuracy
LBP	40.02%

MLBP	43.75%
HOG	44.12%
SIFT	42.26%
Multi-scale version of SIFT {1,3,5,7}	44.19%
SIFT-Rank [40]	42.37%
Bio-inspired features	38.72%
Our Proposed Feature Descriptor	47.87%

Overall benchmark comparisons. In this experiment, we compared our approach to state-of-the-art AIFR approaches. Comparative results are reported in Table 2. The algorithms in Table 2 were tuned to the best settings according to their original papers, and all methods used the same training (10,000 training subjects) and testing datasets (10,000 testing subjects different from the training subjects). From the results in Table 2, we have several observations. First, by applying the proposed IFA matching framework on the HOG features, we can obtain a better result than the method (applying HFA matching framework on the HOG features) in [41]. This shows the superiority of IFA over HFA. Second, it can be seen that our approach (applying IFA on the MEFA features) delivers significant improvements in recognition accuracy. Recognition performance was further improved by extending our matching framework to combine multiple local features, such as MEFA + SIFT (or HOG). Especially, by applying our matching framework to the combined features (MEFA + SIFT + MLBP), 94.59% recognition accuracy was obtained on MORPH Album 2. This is an extremely encouraging result considering the complexity of this dataset.

Table 2. Comparison of our approach with state-of-the-art approaches on the MORPH Album 2 dataset.

Algorithms	Recognition Accuracies
FaceVACS [31]	78.90%
Park et al. (2010) [18]	79.80%
Du et al. (2012) [33]	79.24%
Li et al. (2011) [20]	83.90%
Klare et al. (2011) [34]	79.08%
Otto et al. (2012) [35]	81.27%
Zhen et al. (2013) [37]	86.12%
Gong et al. (2014) [41]	91.14%
IFA matching framework	**92.26%**
on the HOG features	
Our approach (MEFA only)	**93.80%**
Our approach (MEFA+SIFT)	**94.16%**
Our approach (MEFA+MLBP)	**94.22%**
Our approach (MEFA+MLBP+SIFT)	**94.59%**

3.2. Experiments on the FGNET dataset

In this experiment, we compared our approach with state-of-the-art results on another public-domain face aging dataset, FGNET [36]. Note that while the number of the subjects in this dataset is relatively small (1002 face images from 82 people), FGNET has more images per subject than MORPH. FGNET also suffers from the fact that, in addition to age variations, there are large pose, lighting, and expression variations. Following the training and testing split scheme used in [20][41], we also use the leave-one-out scheme for performance evaluation. Comparative results are reported in Table 3. Our approach significantly outperforms the others.

Table 3. Comparison of our approach with state-of-the-art results on FGNET. The rank-1 identification rates are listed.

Algorithms	Recognition Accuracies
FaceVACS [31]	26.4%
Park et al. (2010) [18]	37.4%
Li et al. (2011) [20]	47.5%
Gong et al. (2013) [41]	69.0%
Our approach	**76.2%**

3.3. Experiments on LFW

As discussed above, our approach can easily be extended to address the general face recognition problem. In order to verify the generalizability of our new approach, we also performed experiments on the LFW database [50], a well-used database containing 13,233 face images from 5,749 different subjects that simulates real-life since the faces are from the TV news. Here, we strictly followed the restricted setting (no outside training data were used, even for landmark detection), and measure the performance of the proposed approach by performing the 10 fold cross validation. Experiments are strictly independent for each

fold, and the averaged results are reported. Comparative results are reported in Table 4. One thing to note is that our approach is originally developed for age invariant face recognition task. Directly applying it on the LFW database (which is not a face aging database) cannot fully demonstrate its advantages. Nevertheless, our approach still obtains a good result on the LFW database, only slightly lower than [49] (88.86% versus 88.97%). This demonstrates the excellent generalizability of our approach.

Table 4. Verification performance comparison on the LFW dataset (under the restricted setting).

Approach	Verification Accuracy
Nowak [43]	0.7393
Hybrid descriptor-based [44]	0.7847
V1-like/MKL [45]	0.7935
APEM (fusion) [46]	0.8408
MRF-MLBP [47]	0.7908
Fisher vector faces [48]	0.8747
Eigen-PEP [49]	0.8897
Our approach	**0.8886**

4. Conclusions

Here, we propose the novel maximum entropy feature descriptor (MEFD) for AIFR. By maximizing the code entropy, we show that face recognition performance can be improved by using a more compact and discriminative feature descriptor. In addition, we present a new feature-matching framework called identity factor analysis (IFA) to further improve recognition performance. Extensive experiments on several public-domain face datasets (MORPH, FGNET, and LFW) clearly show the effectiveness and generalizability of our new approach.

5. Acknowledgments

This work was supported by grants from National Natural Science Foundation of China (61103164 and 61125106),
Natural Science Foundation of Guangdong Province (2014A030313688), Guangdong Innovative Research Team Program (No.201001D0104648280), Australian Research Council Projects (DP-120103730, DP-140102164, FT-130101457, and LP-140100569), the National Basic Research Program of China (2015CB352501), and the Key Research Program of the Chinese Academy of Sciences (Grant No. KGZD-EW-T03).

References

[1] Andreas Lanitis, "A survey of the effects of aging on biometric identity verification," Int. J. Biometrics, vol. 2, no. 1, pp. 34– 52, Dec. 2010.

[2] Narayanan Ramanathan, Rama Chellappa, and Soma Biswas, "Computational methods for modeling facial aging: A survey," J. Vis. Lang. Comput., vol. 20, no. 3, pp. 131–144, 2009.

[3] Yun Fu and Thomas S. Huang, "Human age estimation with regression on discriminative aging manifold," IEEE Transactions on Multimedia, vol. 10, no. 4, pp. 578–584, 2008.

[4] Xin Geng, Zhi-Hua Zhou, and Kate Smith-Miles, "Automatic age estimation based on facial aging patterns," IEEE Trans. Pattern Anal. Mach. Intell., vol. 29, no. 12, pp. 2234–2240, 2007.

[5] Guodong Guo, Yun Fu, Charles R. Dyer, and Thomas S. Huang, "Image-based human age estimation by manifold learning and locally adjusted robust regression," IEEE Transactions on Image Processing, vol. 17, no. 7, pp. 1178–1188, 2008.

[6] Guodong Guo, Guowang Mu, Yun Fu, and Thomas S. Huang, "Human age estimation using bio-inspired features," in CVPR, 2009, pp. 112–119.

[7] Young H. Kwon and Niels Da Vitoria Lobo, "Age classification from facial images," in In Proc. IEEE Conf. Computer Vision and Pattern Recognition, 1999, pp. 762–767.

[8] A. Lanitis, C. Draganova, and C. Christodoulou, "Comparing different classifiers for automatic age estimation," IEEE Trans. Syst., Man, Cybern, vol. 34, no. 1, pp., 621-628, Feb. 2004.

[9] Albert Montillo and Haibin Ling, "Age regression from faces using random forests," in ICIP, 2009, pp. 2465–2468.

[10] Narayanan Ramanathan and Rama Chellappa, "Face verification across age progression," IEEE Transactions on Image Processing, vol. 15, no. 11, pp. 3349–3361, 2006.

[11] Junyan Wang, Yan Shang, Guangda Su, and Xinggang Lin, "Age simulation for face recognition," in ICPR (3), 2006, pp. 913–916.

[12] Shuicheng Yan, Huan Wang, Xiaoou Tang, and Thomas S. Huang, "Learning auto-structured regressor from uncertain nonnegative labels," in ICCV, 2007, pp. 1–8.

[13] Shaohua Kevin Zhou, Bogdan Georgescu, Xiang Sean Zhou, and Dorin Comaniciu, "Image based regression using boosting method," in ICCV, 2005, pp. 541–548.

[14] Andreas Lanitis, Christopher J. Taylor, and Timothy F. Cootes, "Toward automatic simulation of aging effects on face images," IEEE Trans. Pattern Anal. Mach. Intell., vol. 24, no. 4, pp. 442–455, 2002.

[15] Jin-Li Suo, Song Chun Zhu, Shiguang Shan, and Xilin Chen, "A compositional and dynamic model for face aging," IEEE Trans. Pattern Anal. Mach. Intell., vol. 32, no. 3, pp. 385–401, 2010.

[16] Jin-Li Suo, Xilin Chen, Shiguang Shan, andWen Gao, "Learning long term face aging patterns from partially dense aging databases," in ICCV, 2009, pp. 622–629.

[17] Norimichi Tsumura, Nobutoshi Ojima, Kayoko Sato, Mitsuhiro Shiraishi, Hideto Shimizu, Hirohide Nabeshima, Syuuichi Akazaki, Kimihiko Hori, and Yoichi Miyake, "Image based skin color and texture analysis/synthesis by extracting hemoglobin and melanin information in the skin," ACM Trans. Graph., vol. 22, no. 3, pp. 770–779, 2003.

[18] Unsang Park, Yiying Tong, and Anil K. Jain, "Age-invariant face recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol. 32, no. 5, pp. 947–954, 2010.
[19] Haibin Ling, Stefano Soatto, Narayanan Ramanathan, and David W. Jacobs, "Face verification across age progression using discriminative methods," IEEE Transactions on Information Forensics and Security, vol. 5, no. 1, pp. 82–91, 2010.
[20] Zhifeng Li, Unsang Park, and Anil K. Jain, "A discriminative model for age invariant face recognition," IEEE Transactions on Information Forensics and Security, vol. 6, no. 3-2, pp. 1028–1037, 2011.
[21] Timo Ojala, Matti Pietik¨ainen, and Topi M¨aenp¨a¨a, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," IEEE Trans. Pattern Anal. Mach. Intell., vol. 24, no. 7, pp. 971–987, 2002.
[22] Krystian Mikolajczyk and Cordelia Schmid, "A performance evaluation of local descriptors," IEEE Trans. Pattern Anal. Mach. Intell., vol. 27, no. 10, pp. 1615–1630, 2005.
[23] D. Lowe, "Distinctive image features from scale-invariant keypoints," Int'l Journal of Computer Vision, vol. 60, no. 2, pp. 91-110, 2004.
[24] Cong Geng and Xudong Jiang, "Face recognition using sift features," in ICIP, 2009, pp. 3313–3316.
[25] Timo Ahonen, Abdenour Hadid, and Matti Pietik¨ainen, "Face recognition with local binary patterns," in ECCV (1), 2004, pp. 469–481.
[26] Timo Ahonen, Abdenour Hadid, and Matti Pietik¨ainen, "Face description with local binary patterns: Application to face recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol. 28, no. 12, pp. 2037–2041, 2006.
[27] Peter N. Belhumeur, Jo ˜a o P. Hespanha, and David J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," IEEE Trans. Pattern Anal. Mach. Intell., vol. 19, no. 7, pp. 711–720, 1997.
[28] Brendan Klare, Zhifeng Li, and Anil K. Jain, "Matching forensic sketches to mug shot photos," IEEE Trans. Pattern Anal. Mach. Intell., vol. 33, no. 3, pp. 639–646, 2011.
[29] Xiaogang Wang and Xiaoou Tang, "A unified framework for subspace face recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol. 26, no. 9, pp. 1222–1228, 2004.
[30] Karl Ricanek Jr. and Tamirat Tesafaye, "Morph: A longitudinal image database of normal adult age-progression," in FG, 2006, pp. 341–345.
[31] FaceVACS Software Developer Kit, Cognitec Systems GbmH, http://www.cognitec-systems.de.
[32] Xiaogang Wang, Xiaoout Tang, "Random sampling LDA for face recognition," in CVPR, 2004, pp. 259–265.
[33] J. Du, C. Zhai, and Y. Ye, "Face aging simulation and recognition based on NMF algorithm with sparseness constraints," *Neurocomputing*, 2012.
[34] B. Klare and A. K. Jain, "Face Recognition Across Time Lapse: On Learning Feature Subspaces", *IJCB*, Washington, DC, Oct. 11-13, 2011.
[35] C. Otto, H. Han, and A. K. Jain, "How Does Aging Affect Facial Components", *ECCV WIAF Workshop*, Florence, Italy, Oct. 7-13, 2012.
[36] FG-NET Aging Database, http://www.fgnet.rsunit.com/.
[37] L. Zhen and P. Matti, and L. Stan, "Learning Discriminant Face Descriptor", *PAMI* 2013, pp. 289-302.
[38] W. Zhang, X. Wang and X. Tang, "Coupled information-theoretic encoding for face photo-sketch recognition", *CVPR* 2011, pp. 513-520.
[39] Shan Caifeng and Gritti Tommaso, Learning Discriminative LBP-Histogram Bins for Facial Expression Recognition, *BMVC*, 2008.
[40] Toews, M., Wells, W., SIFT-Rank: "Ordinal description for invariant feature correspondence", *CVPR* 2009, pp. 172 – 177.
[41] D. Gong, Z. Li, D. Lin, J. Liu, X. Tang, "Hidden Factor Analysis for Age Invariant Face Recognition", ICCV 2013.
[42] B. Chen, C. Chen, and W. H. Hsu, "Cross-Age Reference Coding for Age-Invariant Face Recognition and Retrieval", ECCV 2014.
[43] E. Nowak and F. Jurie, "Learning visual similarity measures for comparing never seen objects," In *Proc. Computer Vision and Pattern Recognition*, pp. 1–8, 2007.
[44] L. Wolf, T. Hassner, Y. Taigman, "Descriptor based methods in the wild," In *Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition*, 2008.
[45] N. Pinto, J.J. DiCarlo, D.D. Cox, "How far can you get with a modern face recognition test set using only simple features?" In *Computer Vision and Pattern Recognition,* 2009.
[46] H. Li, G. Hua, Z. Lin, J. Brandt, and J. Yang, "Probabilistic elastic matching for pose variant face verification," In *CVPR*, 2013.
[47] S.R. Arashloo and J. Kittler, "Efficient processing of mrfs for unconstrained-pose face recognition," In *BTAS*, 2013.
[48] K. Simonyan, O.M. Parkhi, A. Vedaldi, and A. Zisserman, "Fisher vector faces in the wild," In *BMVC*, 2013.
[49] Haoxiang Li, Gang Hua, Xiaohui Shen, Zhe Lin, and Jonathan Brandt, "Eigen-PEP for Video Face Recognition," *Asian Conference on Computer Vision (ACCV)*, 2014.
[50] G.B. Huang, M. Mattar, T. Berg, E. Learned-Miller, and A. Hanson, "Labelled faces in the wild: A database for studying face recognition in unconstrained environments," *Technical Report 07-49*, University of Massachusetts, Amherst, October 2007.
[51] X. Tang, "Texture information in ren-length matrices," *IEEE Transactions on Image Processing*, 1998.
[52] X. Tang and Z. Li, "Audio-guided video based face recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, 2009.
[53] Zhifeng Li, Dahua Lin, and Xiaoou Tang, "Nonparametric Discriminant Analysis for Face Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Volume 31, Issue 4, 2009.