

Dual-reference Face Retrieval: What Does He/She Look Like at Age ‘X’?

BingZhang Hu
University of East Anglia
bingzhang.hu@uea.ac.uk

Feng Zheng
University of Sheffield
cip12fz@sheffield.ac.uk

Ling Shao
University of East Anglia
ling.shao@uea.ac.uk

Abstract

Face retrieval has received much attention over the past few decades, and many efforts have been made in retrieving face images against pose, illumination, and expression variations. However, the conventional works fail to meet the requirements of a potential and novel task — retrieving a person’s face image at a given age, i.e. ‘what does a person look like at age X ?’ The reason that previous works struggle is that text-based approaches generally suffer from insufficient age labels and content-based methods typically take a single input image as query, which can only indicate either the identity or the age. To tackle this problem, we propose a dual reference face retrieval framework in this paper, where the identity and the age are reflected by two reference images respectively. In our framework, the raw images are first projected on a joint manifold, which preserves both the age and identity locality. Then two similarity metrics of age and identity are exploited and optimized by utilizing our proposed quartet-based model. The quartet-based model is novel as it simultaneously describes the similarity in two aspects: identity and age. The experiment shows a promising result, outperforming hierarchical methods. It is also shown that the learned joint manifold is a powerful representation of the human face.

1. Introduction

Over the past few decades, face retrieval has received great interest in the research community for its potential applications such as finding missing persons[19] and matching criminals with CCTV footage for law enforcement [35]. [39]Face retrieval was first introduced by [39], and has since been a burgeoning area in computer vision. [36] used eigenfaces for face recognition to yield a satisfying result on their own dataset. With more digital image and video data uploaded to the internet than ever before, the requirements of face retrieval tasks become more critical, especially as this data is generally unconstrained in pose, illumination and expression. As a matter of fact, aging is also a paramount factor that compromises the accuracy of face

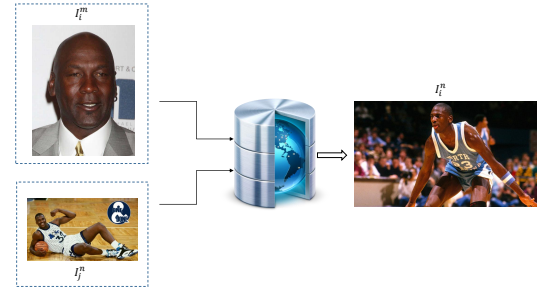


Figure 1. An example of the proposed problem: ‘What does Michael Jordan look like at age X ?’ In this example, we only have a Jordan’s image I_{Jordan}^m at age m in hand and want to retrieve the Jordan’s image I_{Jordan}^n at age n , where n is not a certain numeral and can only be reflected by another image, for example, Shaquille O’Neal’s image $I_{\text{O’Neal}}^n$.

retrieval. However, there are very few approaches which incorporate techniques to deal with age discrepancy for face retrieval tasks. To the best of our knowledge, there are only two works[5, 11] that tackle cross-age face retrieval. [5] proposed a cross-age reference coding method which encodes the low-level features of a face image within an age-invariant reference space. [11] employed an alternating greedy coordinate descent learning algorithm to select age-sensitive and identity-sensitive features to enhance the cross-age face retrieval.

Beyond the conventional content-based face retrieval problems, in this paper, we focus on a much more challenging task — retrieving a person’s face image at a given age. The traditional content-based framework which takes a single query image does not succeed at this task as this image can indicate the identity or the age, but not both. Adding a text query to address the objective age is not a solution either, as it is infeasible to have sufficient age labels for large-scale online datasets. Additionally, it is sometimes not easy to describe an age by a certain numeral and we may just want to look for an image of person A at a similar age stage shown in person B ’s image. Figure. 1 shows an example of the proposed problem. To address these issues, we pro-

pose a novel face retrieval framework with two input reference images, with one indicates the identity and the other denotes the age, which we refer to as dual-reference face retrieval (DRFR).

In our proposed dual-reference face retrieval framework, the raw images are first projected onto a joint manifold, which preserves both the age and identity locality. Subsequently, as the age and identity are measured differently on the joint manifold, a similarity metric for each is exploited and optimized via our proposed quartet-based model. In detail, each quartet sample contains four images, comprising of two people at two different ages. Within each sample, a rectangle can be formed by placing the images of the same identity in the same row and the images with subjects of the same age in the same column. Thus the margin between the hypotenuse and the cathetus of the rectangle can be maximized. It is important to note that the quartet-based model is novel as it simultaneously describes the similarity in two aspects: identity and age.

The contributions of this paper mainly lie in the following three aspects:

- 1) We propose a novel and widely applicable face retrieval task.
- 2) A joint manifold of identity and age is exploited to preserve the localities of these two aspects. A novel quartet-based model coordinated with two Mahalanobis distances is proposed to describe the similarity between image pairs.
- 3) Our proposed DRFR task can be abstracted to a high-level task — dual reference/query retrieval, which might lead to a new research direction.

The remainder of the paper is organized as follows: we review related works in Section 2; our proposal is outlined in detail in Section 3; in Section 4, we discuss the experiments and results; we provide a short conclusion in Section 5.

2. Related Works

As DRFR is newly proposed and to the best of our knowledge, there are no similar works in the literature, we review related works in the areas of face retrieval and age estimation, focusing on those papers which explore facial feature representation, age variation capturing and similarity metric learning.

Facial Feature Representation. A broad array of research has been completed on facial feature representation. As facial features are not the core part of our framework, we just give a rough review here. For a comprehensive review, we refer our readers to [3]. Early works mainly take heuristic features such as LBP[1], SIFT[20], HOG[9], Gabor[24]

or their extensions, LTP[34], FPLBP[38], HDLBP[6]. However, designing hand-crafted features is a trial and error process which is less than adequate for our purpose. Another branch of research regarding facial features is based on utilizing deep learning. [18] proposed unsupervised algorithms to learn feature representation automatically and [10] employ deep learning to learn similarity scores.

Age Variation Capturing. Age variation capturing is rarely considered in conventional face retrieval approaches because in most works to date, features are required to be age-invariant. In contrast, as we need to retrieve the image of a 'certain' age, we need a framework which embeds the age variation in our final facial representations. Approaches capturing age variation can primarily be found in age estimation literature. The earliest approach of age estimation based on facial images dates back to 1994, [21] uses geometric features, in which the ratios between different measurements of facial landmarks (*e.g.* eyes, chin, nose, mouth, etc.) are calculated to classify the individual into three age groups, namely *infants*, *young adults* and *senior adults*. Unfortunately, it suffers in distinguishing young and old adults as both the shape and texture of the face change during aging [33]. To overcome the drawbacks of geometric features, the Active Appearance Model(AAM) is proposed in [7]. AAM is able to simultaneously capture the shape and texture information of face images. Our proposal is inspired by Ageing Pattern Subspace [13], in which a serial of a person's images is treated as an aging pattern. However, our proposed joint manifold is very different because we also embed the identity information at the same time.

Ranking Model. Once the proper facial image representation is selected, ranking the similarities is the next concern. Conventional ranking models can be divided into three general approaches: point-wise[2], pair-wise[8] and list-wise[4]. Pair-wise ranking models are preferred in most retrieval works as they take into account the correlations between each sample and have acceptable computation complexity. Triplet-based models are an extension of pair-wise ranking models, and are discussed in [32]. The margin between the matched pairs and the mismatched pairs in each triplet sample is maximized to learn the ranking function. Compared to traditional pair-wise methods, triplet-based model considers not only the inter-class variations, but also the intra-class variations. As both the identity and age similarities need to be ranked in our task, which the triplet-based model is not capable of, we further extend it to a quartet-based model.

3. Dual-Reference Face Retrieval

For convenience, define I_i^m as an image of the individual with identity i at age m . Input an image pair (I_i^m, I_j^n) , where i is the target identity and n is the objective age, thus our required output is I_i^n . As discussed, DRFR consists of

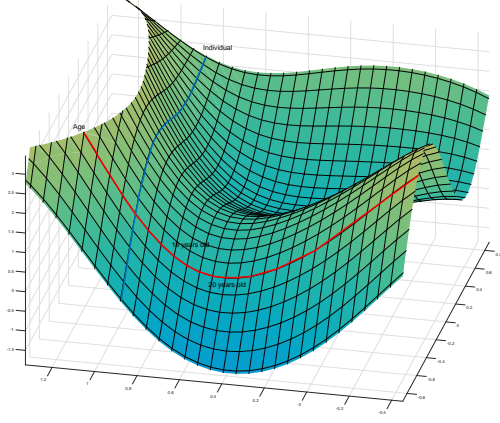


Figure 2. An illustration of the joint manifold of age and identity. The red line indicates the age manifold, while the blue line stands for the individual manifold. The localities of age and identity are preserved in the joint manifold.

two stages. Firstly, a mapping function is learned to project the raw images onto a joint manifold. Subsequently, to measure the similarity between each pair of images, the two metrics are learned on the low-dimensional space, based on a quartet model. We devote the rest of this section to outlining these two stages.

3.1. Joint Manifold

A face image with \mathcal{D} -dimensional feature representation can be considered as a point in the \mathcal{D} -dimensional space containing rich information such as age, gender, race, identity. Manifold learning is first proposed in [29], in which they believe that the high-dimensional data is sampled from a smooth low-dimensional manifold. Thus it is natural that information from a facial image can be represented within low-dimensional manifolds embedded in a high-dimensional image space [12, 27, 17]. Many applications already utilize low-dimensional manifolds to embed human face images, such as face recognition [17] and age estimation [14]. However, our proposed joint manifold as illustrated in Figure. 2 is very different; instead of treating the age and identity as two separate degrees of freedom in a single manifold, we make an assumption that the age and identity are both manifolds sampled from a higher-dimensional manifold. This assumption is proved in Section 4.

Let \mathcal{X} be the original representation of the raw images and \mathcal{Y} be the low-dimensional joint manifold, define the mapping function of the joint manifold to be $f : \mathcal{X} \rightarrow \mathcal{Y}$. Since both the locality of the age and identity can be represented as matrices, let S denote the set of all such similarity matrices. Specifically, the matrix $S^n \in S$ reflects the simi-

larity among all the individuals' images at age n ; similarly, S_i denotes the similarity over those images belonging to an individual with identity i across all ages. The desired properties of f are discussed below.

3.1.1 Preserving locality of individual space

We first calculate the similarity matrix S^n . In detail, among all the images at age n , if two images are nearby in original feature space \mathcal{X} , we mark the similarity as $\exp\left(-\frac{\|x_i^n - x_j^n\|_2^2}{t}\right)$, where $x_i^n \in \mathcal{X}$ is the original feature representation of image I_i^n and $\|\cdot\|_2^2$ is the l_2 -norm, otherwise their similarity is 0. Thus the similarity matrix S^n under age n is calculated as:

$$S^n(x_i^n, x_j^n) = \begin{cases} \exp\left(-\frac{\|x_i^n - x_j^n\|_2^2}{t}\right) & \text{if } x_j^n \in \mathcal{N}(x_i^n), \\ 0 & \text{otherwise,} \end{cases} \quad (1)$$

where $\mathcal{N}(x_i^n)$ denotes the neighbors of x_i^n . To preserve the locality, we require the nearby points in \mathcal{X} to remain close to each other after being embedded into $\mathcal{Y} = f(\mathcal{X})$, thus we optimize the function:

$$\min_f \sum_n \sum_{i,j} \|f(x_i^n) - f(x_j^n)\|_2^2 S^n(x_i^n, x_j^n). \quad (2)$$

3.1.2 Preserving locality of age space

Similarly, to calculate the age similarity matrix S_i , we gather all the images of the individual i , and assign $\exp\left(-\frac{\|x_i^n - x_i^m\|_2^2}{t}\right)$ as the similarity if $m - n$ is below a threshold ε , otherwise the similarity is 0:

$$S_i(x_i^n, x_i^m) = \begin{cases} \exp\left(-\frac{\|x_i^n - x_i^m\|_2^2}{t}\right) & \text{if } |m - n| < \varepsilon, \\ 0 & \text{otherwise.} \end{cases} \quad (3)$$

To preserve the local smoothness, we optimize the function:

$$\min_f \sum_i \sum_{m,n} \|f(x_i^n) - f(x_i^m)\|_2^2 S_i(x_i^n, x_i^m). \quad (4)$$

3.2. Similarity Metric Learning Based on a Quartet Model

After both the original age and identity spaces are mapped onto a joint manifold, different measurements should be taken to obtain the similarity of the two aspects. In this paper, two similarity metrics are learned based on a novel quartet model, which is a graph with 4 vertices as shown in Figure. 3. The vertices sets $V = \{f(x_i^m), f(x_i^n), f(x_j^m), f(x_j^n)\}$ are the embedded points of $\{(x_i^m), (x_i^n), (x_j^m), (x_j^n)\}$, and the edges are defined as the distance between each embedded point. We use $\Phi(\cdot, \cdot)$ to

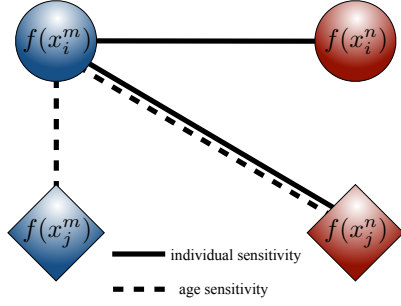


Figure 3. An illustration of our proposed quartet model. Two types of distances are considered: dashed lines refer to age sensitivity while solid lines refer to individual sensitivity.

denote the difference measurement function whereby the smaller $\Phi(\cdot, \cdot)$ is, the more similar the two images are. In the following of this subsection, the properties of the desired metrics are introduced.

3.2.1 Individual metric

Considering two image pairs (x_i^m, x_i^n) and (x_j^m, x_j^n) , which are shown in the quartet model in Figure. 3, it is very clear that on the individual metric, the distance between x_i^m and x_i^n is smaller than that between x_i^m and x_j^n , because these two pairs of images both have the age gap $m - n$ while the first image pair (x_i^m, x_i^n) belongs to the same individual i . Mathematically, there is:

$$\Phi_{ind}(f(x_i^m), f(x_i^n)) < \Phi_{ind}(f(x_i^m), f(x_j^n)) \quad (5)$$

$$\forall(i, j, m, n),$$

where Φ_{ind} measures the individual difference between any pair of images.

Additionally, the distances between image pair (x_i^m, x_j^m) and (x_i^m, x_j^n) are supposed to be similar because the individual metric is uncorrelated with the age, which can be written as:

$$\Phi_{ind}(f(x_i^m), f(x_j^m)) = \Phi_{ind}(f(x_i^m), f(x_j^n)) \quad (6)$$

$$\forall(i, j, m, n),$$

3.2.2 Age metric

Similarly on the age metric, the distance between image pair (x_i^m, x_j^m) is smaller than that between (x_i^m, x_j^n) , and the distances are close if the age gap within each image pair is same. Thus we have:

$$\Phi_{age}(f(x_i^m), f(x_j^m)) < \Phi_{age}(f(x_i^m), f(x_j^n)) \quad (7)$$

$$\forall(i, j, m, n),$$

$$\Phi_{age}(f(x_i^m), f(x_j^n)) = \Phi_{age}(f(x_i^m), f(x_i^n)) \quad (8)$$

$$\forall(i, j, m, n),$$

where Φ_{age} measures the age difference between any pair of images.

3.2.3 Quartet loss

To obtain the discussed characteristics of the individual and age metrics, a loss function which maximize the margin between the distances in Eq. 5 and Eq. 7, and meanwhile minimize the margin between the distances in Eq. 6 and Eq. 8 is designed. For convenience, we first define d as the distance of two images embedded in the joint manifold \mathcal{Y} : $d_{ij}^{mn} = f(x_i^m) - f(x_j^n)$ and take the Mahalanobis distance as the distance measurement. Thus the $\Phi(\cdot, \cdot)$ can be written as:

$$\Phi_{age}(f(x_i^m), f(x_j^n)) = d_{ij}^{mn \top} \mathbf{M}_{age} d_{ij}^{mn}, \quad (9)$$

$$\Phi_{ind}(f(x_i^m), f(x_j^n)) = d_{ij}^{mn \top} \mathbf{M}_{ind} d_{ij}^{mn},$$

where \mathbf{M}_{age} and \mathbf{M}_{ind} are the Mahalanobis matrices. To maximize the margin, the hinge loss function:

$$H(y) = \max(0, \delta - y) \quad (10)$$

is employed.

Thereby for a quartet sample indexed by (i, j, m, n) , the loss \mathcal{L}_{ij}^{mn} can be defined as:

$$\begin{aligned} \mathcal{L}_{ij}^{mn} = & H(d_{ij}^{mn \top} \mathbf{M}_{age} d_{ij}^{mn} - d_{ij}^{mm \top} \mathbf{M}_{age} d_{ij}^{mm}) \\ & + H(d_{ij}^{mn \top} \mathbf{M}_{ind} d_{ij}^{mn} - d_{ii}^{mn \top} \mathbf{M}_{ind} d_{ii}^{mn}) \\ & + \|d_{ij}^{mn \top} \mathbf{M}_{age} d_{ij}^{mn} - d_{ii}^{mn \top} \mathbf{M}_{age} d_{ii}^{mn}\|_2^2 \\ & + \|d_{ij}^{mn \top} \mathbf{M}_{ind} d_{ij}^{mn} - d_{ij}^{mm \top} \mathbf{M}_{ind} d_{ij}^{mm}\|_2^2. \end{aligned}$$

And the loss over the whole training set is

$$\mathcal{L} = \sum_{i,j,m,n} L_{ij}^{mn}. \quad (11)$$

3.3. Optimization

Considering the loss function \mathcal{L} and the joint manifold as the regularization term, the overall objective function is:

$$\begin{aligned} \mathcal{J} = & \mathcal{L} + \sum_n \sum_{i,j} \|d_{ij}^{nn}\|_2^2 S^n(x_i^n, x_j^n) \\ & + \sum_i \sum_{m,n} \|d_{ii}^{mn}\|_2^2 S_i(x_i^m, x_i^n), \quad (12) \end{aligned}$$

$$\text{s.t. } \mathbf{M}_{ind} \succeq 0, \mathbf{M}_{age} \succeq 0,$$

where $\mathbf{M} \succeq 0$ implies that \mathbf{M} is a semi-definite positive matrix, thus pseudometrics are allowed.

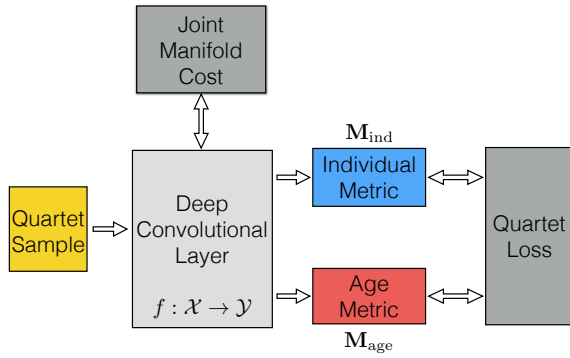


Figure 4. The architecture of our proposed deep network. It takes quartet samples as input, and the deep convolutional layer is trained to explore the joint manifold embedding. Subsequently, the distances between embedded images are measured by two independent metrics – individual metric and age metric. Finally the distances are feed to the last layer to optimize the quartet loss.

As both the Mahalanobis matrices \mathbf{M}_{age} and \mathbf{M}_{ind} as well as the embedding function f need to be learned in Eq. 12, we employ a deep network to optimize them jointly. The architecture of the proposed network is discussed in the following sections.

3.3.1 Deep network architecture

Our quartet-based network architecture is shown in Figure. 4, which jointly optimize the manifold embedding function f and two Mahalanobis matrices. This network takes quartet samples as input. Each quartet sample contains an image set $Q = \{x_i^m, x_i^n, x_j^m, x_j^n\}$, which are the images of the person i and j at his m and n age stage. The images are firstly passed through a convolutional network with a deep architecture, which can extract prolific and robust age and identity information from a facial image while preserving the locality. The deep convolutional network takes the joint manifold cost as the loss function. Subsequently, the distance between the outputs of the deep architecture, for example, $f(x_i^m)$ and $f(x_i^n)$ are measured via two independent metrics, which are namely, age metric and individual metric. With the distances between each image pairs, the quartet loss are thus optimized and the gradients are back-propagated to update the \mathbf{M} .

Deep convolutional layer. In our model, the deep convolution layer is trained to explore the joint manifold of the age and identity. As discussed in the Section 3.1, the joint manifold is supposed to keep the locality structure, thus the Eq. 2 and Eq. 4 are taken as the joint manifold cost. In the experiment, we first compute the similarity matrix across

Table 1. Architecture of proposed deep network

layer	input	kernel	output
conv1	$220 \times 220 \times 3$	$7 \times 7 \times 3$	$110 \times 110 \times 64$
pool1	$110 \times 110 \times 64$	$3 \times 3 \times 64$	$55 \times 55 \times 64$
rnorm1	$55 \times 55 \times 64$		$55 \times 55 \times 64$
conv2a	$55 \times 55 \times 64$	$1 \times 1 \times 64$	$55 \times 55 \times 64$
conv2	$55 \times 55 \times 64$	$3 \times 3 \times 64$	$55 \times 55 \times 192$
rnorm2	$55 \times 55 \times 192$		$55 \times 55 \times 192$
pool2	$55 \times 55 \times 192$	$3 \times 3 \times 192$	$28 \times 28 \times 192$
conv3a	$28 \times 28 \times 192$	$3 \times 3 \times 192$	$28 \times 28 \times 192$
conv3	$28 \times 28 \times 192$	$3 \times 3 \times 192$	$28 \times 28 \times 384$
pool3	$28 \times 28 \times 384$	$3 \times 3 \times 384$	$14 \times 14 \times 384$
conv4a	$14 \times 14 \times 384$	$1 \times 1 \times 384$	$14 \times 14 \times 384$
conv4	$14 \times 14 \times 384$	$3 \times 3 \times 384$	$14 \times 14 \times 256$
pool4	$14 \times 14 \times 256$	$3 \times 3 \times 256$	$7 \times 7 \times 256$
concat	$7 \times 7 \times 256$		$7 \times 7 \times 256$
fc1	$7 \times 7 \times 256$		$1 \times 32 \times 128$
fc2	$1 \times 32 \times 128$		$1 \times 1 \times 128$

the whole dataset while for each input batch, only the involved locality constrains need to be satisfied during training, which leads to a great computation saving. As a fact, the linear embedding can already reflect the joint manifold, however we employ the deep learning for a better performance. Table. 1 shows the detail of the architecture of the proposed deep convolutional network, which is inspired by the [30].

Individual metric and age metric. At the end of the deep architecture module, the facial images are represented by a d -dimensional feature. To measure the distances between each image, we introduce two Mahalanobis matrices \mathbf{M}_{age} and \mathbf{M}_{ind} . Since Mahalanbis matrices are semi-definite positive, \mathbf{M} can be factorized as $\mathbf{M} = L^T L$. In other words, to learn the individual metric and age metric is equally to learn two projections L_{ind} and L_{age} as:

$$\begin{aligned} \Phi(f(x_i^m), f(x_j^n)) &= d_{ij}^{mn \top} \mathbf{M} d_{ij}^{mn} \\ &= \|L f(x_i^m) - L f(x_j^n)\|_2^2 \end{aligned} \quad (13)$$

In our architecture, the two metrics layer are inner product layers with independent weights. The euclidean distance in the projected space is the corresponding Mahalanbis distance. It is not hard to update the matrix L via the loss function Eq. 12 while how to ensure \mathbf{M} being semi-positive is a problem. Inspired by [31], we take a trick when updating on L happens. After L is updated by the network, we check all the eigenvalue of the matrix L and change the most negative eigenvalue to zero and then update L again to make it closer to a semi-positive matrix.

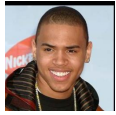


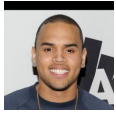


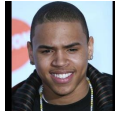
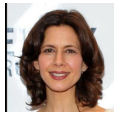



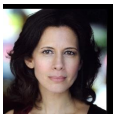




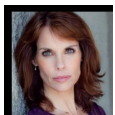
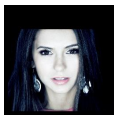



Query Pairs		Retrieval Results				
Identity Reference	Age Reference	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5
 I_{1950}^{16}	 I_{1674}^{25}	 I_{1950}^{24}	 I_{1950}^{22}	 I_{1097}^{39}	 I_{1674}^{20}	 I_{1950}^{17}
 I_{731}^{46}	 I_{1131}^{40}	 I_{731}^{42}	 I_{1131}^{42}	 I_{731}^{48}	 I_{1130}^{40}	 I_{1078}^{40}
 I_{1934}^{16}	 I_{1759}^{26}	 I_{643}^{50}	 I_{1903}^{16}	 I_{1813}^{17}	 I_{1984}^{14}	 I_{1950}^{24}

Figure 5. Experiment results on CADC dataset. The first row and second row are selected two good retrieval results. We can find that the performance of our DRFR is encouraging though the third output in the first retrieval is not very accurate. The third row is a bad retrieval which is totally wrong in top-5 outputs. However, we think the failure in the 3rd example is because the age reference image is miss-labelled and hard to recognize, even to human.

4. Experiment

Due to the proposed problem is very novel and currently not widely studied, there are very limited datasets suitable for the experiment. It is hard to search datasets with large range labelled age and vast identity. And that is also why this paper studies this problem. In the experiment, we evaluate our DRFR on three famous face recognition and age estimation datasets: Cross-Age Celebrity Dataset(CACD)[5], FGNet[22], and MORPH[28]. The statistics of these datasets are shown in Table. 2. We can see that CACD[5] contains the largest number of images while MORPH[28] has the most subjects. Although FGNet[22] is small compared to the other two datasets, its huge age gap makes it a good platform to evaluate our DRFR. Moreover, we evaluate the age distribution of these datasets, shown in Figure. 6.

4.1. Experiment on CACD

Dataset Setting The Cross-Age Celebrity Dataset is collected for the cross age face retrieval task in [5], and it contains 163446 images from 2000 celebrities with the age ranging from 16 to 62. The large scale data with high age variations supplies a high quality experiment environment

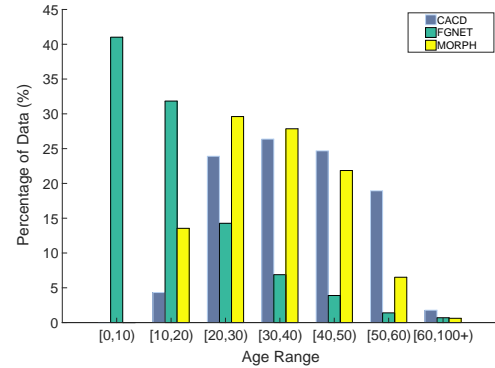


Figure 6. The age range distribution of the experiment datasets.

for our DRFR. However, it is noteworthy that although the age ranges from 16 to 62, the maximum age gap for each celebrity is 9 years old, as all the collected images are taken from 2003 to 2014. In details, the age gaps are stepping at 1 year old from (14-23) to (53-62), thus there are 40 age gaps in total. On average, each age gap contains 4000 images of 50 celebrities. Following the settings in [5], we take 60% data as training data and the remaining data for testing. The training data is picked uniformly from each age gap to

Dataset	Images	Subjects	Images/sub.	Age gap
CACD	163446	2000	81.7	0-9
FGNet	1002	82	12.2	0-45
MORPH	55134	13618	4.1	0-5

Table 2. Statistics of the Datasets

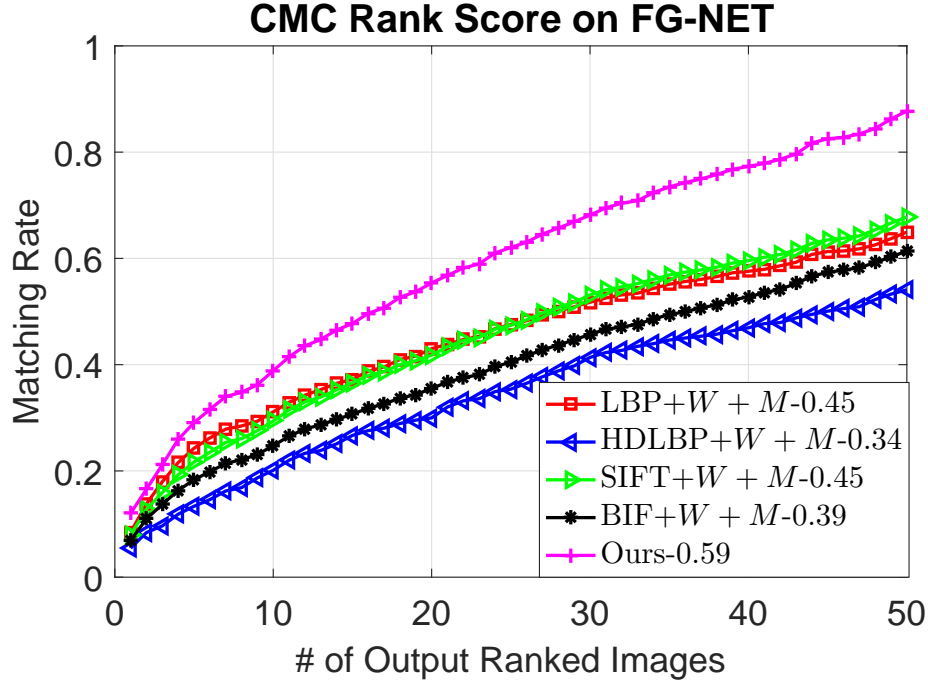


Figure 7. The result of the experiment on FGNet.

ensure all the age gaps are covered. For the testing data, as there are averagely 8 different images for each celebrity at each age, we further split the testing data into 8 subsets for the evaluation. We employed HDLBP[6] as the original features during the training session, and the ε in Eq. 11 was set as 3 empirically to compute the similarity matrix set S . The transform matrix \mathbf{W} of projection function f was initialized randomly with a Gaussian distribution and the two Mahalanobis matrices were initialized as the identity matrix. It is interesting to mention that though the dimension of \mathcal{Y} is a super-parameter, we found the performance was stable when we change it from 10 to 40. Thus in this paper, the dimension of \mathcal{Y} is fixed as 20.

Evaluation Metrics and Comparison As DRFR can be regarded as a fine-grained retrieval, we use the top- K retrieval accuracy[37] as the evaluation metric. Since there are no works on this task in the literature before, we combined the existing face retrieval approaches with the age estimation methods to form a hierarchical framework and made the comparison. In the combined hierarchical framework, the face retrieval was first conducted regarding the

first reference image as query. Subsequently, we estimated the age of the second reference image and the top 100 candidate images from the face retrieval session. Finally these 100 images are ranked according to the estimated ages. We choose eigenfaces [36] and CARC [5] to perform the face retrieval, and SVR [14] and CCA [15] for the age estimation, thus there are 4 combinations in total.

Results and conclusion We conducted DRFR and the 4 hierarchical methods on the 8 testing subsets and compute the average top- K retrieval accuracy. The results are shown in Table. 3. It shows that when the K is small (less than 6), our proposed DRFR outperformed the other 4 three methods. It is interesting to note that when the allowed output image increases, the accuracy of CARC+CCA is slightly higher than ours. The reason is that CARC is a powerful cross-age retrieval method on CADC[5], and in our settings, each subset only contains approximately 10 images for each subject, it is reasonable for a high accuracy if the face retrieval system can retrieve all the images of the correct identity.

Accuracy% @ top- K	$K=1$	$K=2$	$K=5$	$K=8$	$K=10$
eigenfaces+SVR	14.43	17.25	18.5	19.10	19.20
eigenfaces+CCA	14.97	17.73	18.71	19.24	19.35
CARC+SVR	18.34	22.45	24.30	25.70	26.20
CARC+CCA	18.57	22.25	24.50	26.10	26.40
Our DRFR	20.67	23.75	24.90	25.80	26.33

Table 3. Experiment Results on CACD dataset

4.2. Experiment on FGNet

Dataset Setting FGNet dataset consists of 1002 images of 82 subjects in total. As it is tiny while has high age variations, we conduct experiments using different feature on it to evaluate performance of the joint manifold embedding function f of our proposed framework. Similar with the experiment setting on CACD, we split FGNet into training and test set, avoiding the situation that the same subject shows in both sets. The training set contains 60% images while the rest is left for test. We initialized the variables same with the experiment on CACD.

Comparison with linear embedding method We selected four different feature descriptors as our original feature space \mathcal{X} , which includes: LBP[16], BIF[26], SIFT[25] and HDLBP[6] and employed a linear embedding to make the comparison. To learn the linear embedding, we employed PCA as the embedding technique, which denotes as W , and concatenate the embedded points directly to the metrics layer M in our framework.

Results and conclusion Figure. 7 shows the results of our experiments on different features conducted on FGNet. It can be seen that accuracy is very stable for each features and we believe that the DRFR is very robust to the original feature space. It is worthy to note that the deep learned feature yields the highest accuracy. We conclude that the reason is that the model we used is trained on a large-scale online face dataset and the massive size of training data can help to reveal the latent variables in representing a face. Moreover, one task of the multi-task problem solved in [23] is age estimation, thus that the features work well for DRFR is unsurprising.

4.3. Validation on MORPH

The MORPH dataset has 55134 images of 13618 subjects. Though both the images and subjects are in big amount, the number of images for each subject is only 4.1, which is not sufficient to compromise the quartet samples for training. Thereby instead of training a new model, we conduct a cross-validation on MORPH. We first trained our deep network on the CACD dataset and then run it directly on the MORPH dataset, but the results are not satisfying. However the result improves a lot after we fine tuned our network on the FGNet, and the accuracy at top-5 is aver-

agely 28.6% on 50 runs.

5. Conclusion

In this paper, we proposed a dual-reference face retrieval framework, which tackles the problem of retrieving a person's face image at a given age. In the proposed framework, the retrieval is conducted on a joint manifold and based on two similarity metrics, of which one measures the age similarity and the other measures the identity similarity. The joint manifold is revealed by projecting the images into a low-dimensional space while simultaneously preserving the locality of age and identity. The similarity metrics of age and identity are learned via our proposed quartet model, in which the complex relations between samples in the aspect of age and identity are completely reflected.

We have systematically evaluated our approach on CACD, FGNet and MORPH, and the corresponding results show that the proposed approach achieves promising results on this new task and the framework is stable and robust.

References

- [1] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE transactions on pattern analysis and machine intelligence*, 28(12):2037–2041, 2006.
- [2] R. Baeza-Yates, B. Ribeiro-Neto, et al. *Modern information retrieval*, volume 463. ACM press New York, 1999.
- [3] E. Bagherian and R. W. O. Rahmat. Facial feature extraction for face recognition: a review. In *2008 International Symposium on Information Technology*, volume 2, pages 1–9. IEEE, 2008.
- [4] Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, and H. Li. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th international conference on Machine learning*, pages 129–136. ACM, 2007.
- [5] B.-C. Chen, C.-S. Chen, and W. H. Hsu. Cross-age reference coding for age-invariant face recognition and retrieval. In *European Conference on Computer Vision*, pages 768–783. Springer, 2014.
- [6] D. Chen, X. Cao, F. Wen, and J. Sun. Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3025–3032, 2013.
- [7] T. F. Cootes, G. J. Edwards, C. J. Taylor, et al. Active ap-

- pearance models. *IEEE Transactions on pattern analysis and machine intelligence*, 23(6):681–685, 2001.
- [8] K. Crammer, Y. Singer, et al. Pranking with ranking. In *Nips*, volume 1, pages 641–647, 2001.
 - [9] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893. IEEE, 2005.
 - [10] J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, M. Mao, A. Senior, P. Tucker, K. Yang, Q. V. Le, et al. Large scale distributed deep networks. In *Advances in neural information processing systems*, pages 1223–1231, 2012.
 - [11] L. Du and H. Ling. Cross-age face verification by coordinating with cross-face age verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2329–2338, 2015.
 - [12] D. Fidaleo and M. Trivedi. Manifold analysis of facial gestures for face recognition. In *Proceedings of the 2003 ACM SIGMM workshop on Biometrics methods and applications*, pages 65–69. ACM, 2003.
 - [13] X. Geng, K. Smith-Miles, and Z.-H. Zhou. Facial age estimation by nonlinear aging pattern subspace. In *Proceedings of the 16th ACM international conference on Multimedia*, pages 721–724. ACM, 2008.
 - [14] G. Guo, Y. Fu, C. R. Dyer, and T. S. Huang. Image-based human age estimation by manifold learning and locally adjusted robust regression. *IEEE Transactions on Image Processing*, 17(7):1178–1188, 2008.
 - [15] G. Guo and G. Mu. Joint estimation of age, gender and ethnicity: Cca vs. pls. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–6. IEEE, 2013.
 - [16] R. M. Haralick, K. Shanmugam, et al. Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics*, (6):610–621, 1973.
 - [17] X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang. Face recognition using laplacianfaces. *IEEE transactions on pattern analysis and machine intelligence*, 27(3):328–340, 2005.
 - [18] G. B. Huang, H. Lee, and E. Learned-Miller. Learning hierarchical representations for face verification with convolutional deep belief networks. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2518–2525. IEEE, 2012.
 - [19] A. K. Jain, B. Klare, and U. Park. Face matching and retrieval in forensics applications. *IEEE multimedia*, 19(1):20, 2012.
 - [20] D. R. Kisku, A. Rattani, E. Grosso, and M. Tistarelli. Face identification by sift-based complete graph topology. In *Automatic Identification Advanced Technologies, 2007 IEEE Workshop on*, pages 63–68. IEEE, 2007.
 - [21] Y. H. Kwon and N. D. V. Lobo. Age classification from facial images. In *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR'94., 1994 IEEE Computer Society Conference on*, pages 762–767. IEEE, 1994.
 - [22] A. Lanitis and T. Cootes. Fg-net aging data base. *Cyprus College*, 2002.
 - [23] G. Levi and T. Hassner. Age and gender classification using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 34–42, 2015.
 - [24] C. Liu and H. Wechsler. Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. *IEEE Transactions on Image processing*, 11(4):467–476, 2002.
 - [25] D. G. Lowe. Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2, pages 1150–1157. Ieee, 1999.
 - [26] G. Mu, G. Guo, Y. Fu, and T. S. Huang. Human age estimation using bio-inspired features. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 112–119. IEEE, 2009.
 - [27] X. Niyogi. Locality preserving projections. In *Neural information processing systems*, volume 16, page 153. MIT, 2004.
 - [28] K. Ricanek and T. Tesafaye. Morph: A longitudinal image database of normal adult age-progression. In *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, pages 341–345. IEEE, 2006.
 - [29] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.
 - [30] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015.
 - [31] S. Shalev-Shwartz, Y. Singer, and A. Y. Ng. Online and batch learning of pseudo-metrics. In *International Conference on Machine Learning*, 2004.
 - [32] A. W. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on pattern analysis and machine intelligence*, 22(12):1349–1380, 2000.
 - [33] J. Suo, F. Min, S. Zhu, S. Shan, and X. Chen. A multi-resolution dynamic model for face aging simulation. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.
 - [34] X. Tan and B. Triggs. Enhanced local texture feature sets for face recognition under difficult lighting conditions. *IEEE transactions on image processing*, 19(6):1635–1650, 2010.
 - [35] X. Tang and X. Wang. Face photo recognition using sketch. In *Image Processing. 2002. Proceedings. 2002 International Conference on*, volume 1, pages I–257. IEEE, 2002.
 - [36] M. A. Turk and A. P. Pentland. Face recognition using eigenfaces. In *Computer Vision and Pattern Recognition, 1991. Proceedings CVPR'91., IEEE Computer Society Conference on*, pages 586–591. IEEE, 1991.
 - [37] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu. Learning fine-grained image similarity with deep ranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1386–1393, 2014.
 - [38] L. Wolf, T. Hassner, and Y. Taigman. Descriptor based methods in the wild. In *Workshop on faces in 'real-life' images: Detection, alignment, and recognition*, 2008.
 - [39] W.W.Bledsoe. The model method in facial recognition. Panoramic Research Inc., 1966.