Deep Cost-Sensitive and Order-Preserving Feature Learning for Cross-Population Age Estimation

Kai Li[1,2]*, Junliang Xing[1,2]*, Chi Su[3], Weiming Hu[1,2,4], Yundong Zhang[5], Steve Maybank[6]
[1] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
[2] University of Chinese Academy of Sciences [3] KingSoft Ltd.
[4] CAS Center for Excellence in Brain Science and Intelligence Technology
[5] Vimicro Corporation [6] Birkbeck University of London
{kai.li,jlxing,wmhu}@nlpr.ia.ac.cn suchi@kingsoft.com raymond@vimicro.com sjmaybank@dcs.bbk.ac.uk

Abstract

Facial age estimation from a face image is an important yet very challenging task in computer vision, since humans with different races and/or genders, exhibit quite different patterns in their facial aging processes. To deal with the influence of race and gender, previous methods perform age estimation within each population separately. In practice, however, it is often very difficult to collect and label sufficient data for each population. Therefore, it would be helpful to exploit an existing large labeled dataset of one (source) population to improve the age estimation performance on another (target) population with only a small labeled dataset available. In this work, we propose a Deep Cross-Population (DCP) age estimation model to achieve this goal. In particular, our DCP model develops a two-stage training strategy. First, a novel cost-sensitive multi-task loss function is designed to learn transferable aging features by training on the source population. Second, a novel order-preserving pair-wise loss function is designed to align the aging features of the two populations. By doing so, our DCP model can transfer the knowledge encoded in the source population to the target population. Extensive experiments on the two of the largest benchmark datasets show that our DCP model outperforms several strong baseline methods and many state-of-the-art methods.

1. Introduction

Facial age estimation, *i.e.*, automatically predicting the age from a face image, is a very important yet difficult problem in computer vision. It has many applications such as human-computer interaction [8], age-based face retrieval [22], intelligent surveillance [34], and precision advertising [30], *etc.* Despite decades of studies [20, 21, 7, 37, 9, 12, 10, 26, 2, 23, 15, 28, 1, 24, 36, 3], it still remains a very challenging problem due to many varied factors from face pose, expression, race, gender, image illumination, and noise, to name a few [5].

Roughly speaking, the factors that make age estimation difficult can be divided into two groups. The first group of factors comes from the *extrinsic* appearance variations of the face images, *e.g.*, face pose, expression, and image illumination [14]. The other group is determined by *intrinsic* human genes associated with race and gender [12]. A large portion of previous work focusses on the first group of factors, while the other group of intrinsic factors receives relatively little attention.

Since different populations, *e.g.*, African and Caucasian, females and males, exhibit quite different aging patterns, it is very challenging to design an age estimator which can generalize to faces from different populations. Some previous works suggest performing age estimation within each population separately [11, 10, 14]. However, training a separate model for each population also has its own limitations since it is difficult and expensive to collect and label sufficient training data for each population. Based on the above considerations, instead of resorting to labeling more data, it is better to exploit the existing large sized training data of one (source) population to improve the age estimation performance on another (target) population for which only a small sized set of training data is available.

As discussed above, we are interested in this *cross-population* age estimation problem (Figure 1). The setting of this new problem is that a large set of training data is available for the source population but only a small set of training data is available for the target population. The training data of the source population is used to improve the age estimation performance on the target population without collecting more data for it. In this work, we propose

*These authors contributed equally to this study.

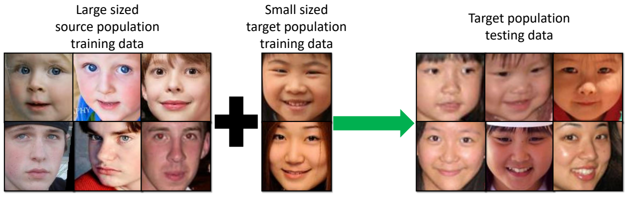

Figure 1. The cross-population age estimation problem. The source and target populations may differ in race and/or gender, and they may have very different aging patterns.

a Deep Cross-Population (DCP) age estimation model to achieve this goal. Instead of manually designing aging features, the DCP age estimation model is based on a Convolutional Neural Network (CNN) which automatically extracts aging features from the input face images. The features are more discriminative and robust to facial appearance variations than the commonly used handcrafted aging features. To obtain high performance on the target population, the DCP model uses a two-stage training strategy:

- In the first stage, age estimation is formulated as a ranking problem because it can take account of the correlations between the age labels. We also design a novel cost-sensitive multi-task loss function for this ranking problem and obtain a model Net^s by training it on the source population. We then create a model Net^t for the target population by copying all the parameters from Net^s. Since CNN can capture useful low-level features independent of the training data [39, 31], the main purpose of this stage is to extract useful and transferable *low-level* aging features from the large sized source population and then transfer them to the target population.

- In the second stage, Net^s and Net^t are fine-tuned. To this end, pair-wise labels (*i.e.*, same age or not) are generated from the source population and the target population. A novel order-preserving pair-wise loss function is designed to bridge the large gaps between aging patterns by aligning the *high-level* aging features of two populations. After this second-stage training, the DCP model effectively transfers the knowledge encoded in the source population to the target population, and thus improve the age estimation performance on the target population even though it has only small sized training data.

To summarize, the main contributions of this work are three-fold:

- We propose a novel Deep Cross Population (DCP) age estimation model. To the best of our knowledge, this DCP model is the first deep model has been designed to solve the challenging cross-population age estimation problem.

- We propose a novel two-stage transfer learning strategy to train this DCP age estimation model with cost-sensitive feature learning and order-preserving feature alignment.

- Our DCP age estimation model exhibits very good performance and outperforms several strong baseline methods as well as many state-of-the-art methods on two of the largest benchmark datasets.

2. Notations and Problem Definition

We first introduce some notation used throughout this paper, and clarify the definition of the problem to solve in this work.

2.1. Notations

We use boldface lowercase letters like $\mathbf{z}$ to denote vectors. The i-th item of $\mathbf{z}$ is denoted as $\mathbf{z}^{(i)}$. Boldface uppercase letters like $\mathbf{Z}$ are used to denote matrices. The transpose of $\mathbf{Z}$ is denoted as $\mathbf{Z}^{\mathrm{T}}$, and the k-th column of $\mathbf{Z}$ is denoted as $\mathbf{Z}^{(k)}$. The notation $\|\cdot\|_{\mathrm{F}}$ is used to denote the Frobenius norm of a vector or matrix, and notation $\mathrm{tr}(\cdot)$ is used for the trace of a matrix.

2.2. Problem definition

In the setting of cross-population age estimation, suppose that there are N^s source population training face images $\mathcal{X}^s = \{\mathbf{X}_i^s, y_i^s\}_{i=1}^{N^s}, y_i^s \in \{1, 2, \ldots, K\}$, where $\mathbf{X}_i^s$ denotes the i-th face image, y_i^s denotes its age label, and K is the total number of different ages. Suppose also that there are N^t target population training face images $\mathcal{X}^t = \{\mathbf{X}_i^t, y_i^t\}_{i=1}^{N^t}, y_i^t \in \{1, 2, \ldots, K\}$. The number of training images from the source population is usually larger than that from the target population, *i.e.*, $N^s > N^t$. Our aim is to train an age estimation model that performs well on testing face images from the target population. It is worth noting that training a good age estimation model usually requires a large amount of training data, whilst only a limited amount of target population training data is available. To overcome this problem, we need to design a new age estimation model which can transfer the knowledge encoded in the source population to the target population in order to obtain satisfactory age estimation performance on the target population.

3. Deep Cross-Population Age Estimation

Since the face images in the target population training set exhibit very different visual patterns in the raw image space to those exhibited by the source population training set, it is not feasible to use directly all the training samples to train a face estimation model for the target population. Inspired by the great success of CNN on learning hierarchical feature representations, our proposed Deep Cross-Population

(DCP) age estimation model deals with this problem by using a new two-stage learning framework, which first learns transferable low-level feature presentation in a novel cost-sensitive feature learning stage, and then learns to align high-level feature presentations across two populations in a novel order-preserving feature alignment stage. In the following, these two stages are explained in detail.

3.1. Cost-sensitive feature learning stage

At this stage a deep model Net^s is trained on the source population training data $\mathcal{X}^s$. The main purpose of this stage is to extract useful and transferable low-level aging features from the large sized source population training data $\mathcal{X}^s$. In order to obtain Net^s, it is necessary to choose the appropriate problem formulation for age estimation, design the network architecture, and design the loss function.

3.1.1 Problem formulation

Age estimation can be naturally formulated as a multi-class classification problem. In this formulation, different ages are assumed to be independent of one another. However, age labels have very strong interrelationships since they form a well-ordered set [24, 36]. On the other hand, regression based methods treat the age labels as numerical values and thus capture the order information for age estimation. However, regression based methods are apt to over-fit the training data as manifested in [2, 27].

In this paper, age estimation is formulated as a ranking problem. There are two main reasons for this choice. First, this ranking formulation is more suitable for characterizing the correlations among different ages [38]. Second, ranking based methods are able to learn more transferable aging features [19] which is desirable for our cross-population age estimation problem. In this ranking based formulation, each age label $y \in \{1, 2, \ldots, K\}$ is treated as a rank. To directly utilize the well-studied classification algorithms, following the reduction framework proposed in [25], the ranking problem is transformed to a series of binary classification problems. In particular, given a training set $\mathcal{X} = \{\mathbf{X}_i, y_i\}_{i=1}^N, y_i \in \{1, 2, \ldots, K\}$. For a given rank (age) k ($1 \leq k < K$), $\mathcal{X}$ is divided into two subsets, $\mathcal{X}_k^+$ and $\mathcal{X}_k^-$, as follows:

$$\begin{cases} \mathcal{X}_k^+ = \{(\mathbf{X}_i, 1) | y_i > k\} \\ \mathcal{X}_k^- = \{(\mathbf{X}_i, 0) | y_i \leq k\}. \end{cases} \tag{1}$$

Next, $\mathcal{X}_k^+$ and $\mathcal{X}_k^-$ are used to train a binary classifier f_k. Since $1 \leq k < K$, $K-1$ binary classifiers $\{f_k\}_{k=1}^{K-1}$ are obtained in total. For a given testing face image $\tilde{\mathbf{X}}$, its age $\tilde{y}$ is predicted by aggregating the $K-1$ decision results as follows,

$$\tilde{y} = 1 + \sum_{k=1}^{K-1} f_k(\tilde{\mathbf{X}}), \tag{2}$$

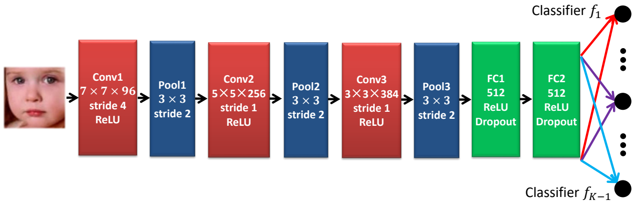

Figure 2. The network architecture for age estimation under the ranking based formulation.

where $f_k(\tilde{\mathbf{X}}) \in \{0, 1\}$ is the classification result of the k-th binary classifier f_k for $\tilde{\mathbf{X}}$.

3.1.2 Network architecture

The architecture of this ranking based age estimation network is shown in Figure 2. There are three convolutional layers, three max pooling layers, and two fully-connected layers. Our choice of this network architecture is motivated by the previous work [23], which employed a similar architecture to perform age group classification and obtained satisfactory performance. It is worth noting that other modern CNN architectures such as ZFNet [40], VGGNet [32] and GoogLeNet [35] can also be used for age estimation, but a comparison of different network architectures is not the focus of this work. The network branches into $K-1$ outputs, where the k-th output corresponds to the binary classifier f_k. Since each binary classification can be treated as one task, we name the network in Figure 2 as Multi-Task Network (MTNet).

3.1.3 Loss function

Given the original training set $\mathcal{X} = \{\mathbf{X}_i, y_i\}_{i=1}^N, y_i \in \{1, 2, \ldots, K\}$, the age label y_i of $\mathbf{X}_i$ corresponds to a vector $\mathbf{y}_i \in \mathbb{R}^{K-1}$ under the ranking based formulation according to Eqn. (1). More specifically, $\mathbf{y}_i$ is defined as follows:

$$\mathbf{y}_i^{(k)} = \begin{cases} 1, & k < y_i \\ 0, & k \geq y_i \end{cases}, \ k \in \{1, \ldots, K-1\}. \tag{3}$$

As a result, the training set now becomes $\mathcal{X} = \{\mathbf{X}_i, \mathbf{y}_i\}_{i=1}^N$. It's worth noting that age estimation is inherently a cost-sensitive problem. For example, when a person's age is y_i, misclassifying y_i as $y_i + 10$ is a more serious mistake than misclassifying y_i as $y_i + 1$. To take this cost sensitivity into consideration, given a training face image $\mathbf{X}_i$ and its age y_i, we use cost$_k(y_i)$ to denote the cost of misclassifying it in the k-th binary classification problem. More specifically, cost$_k(y_i)$ is designed as follows:

$$\text{cost}_k(y_i) = \begin{cases} k - y_i + 1, \ y_i \leq k \\ y_i - k, \ y_i > k \end{cases}, \ k \in \{1, \ldots, K-1\}. \tag{4}$$

We use Θ to collectively denote the parameters of the three convolutional layers. Suppose that there is a total of M convolutional filters in MTNet, then $\Theta = \{\boldsymbol{\theta}_j\}_{j=1}^{M}$, where $\boldsymbol{\theta}_j$ is the vectorization of the j-th filter. Two matrices $\mathbf{W}_{\mathrm{FC1}}$ and $\mathbf{W}_{\mathrm{FC2}}$ are used to denote the parameters of two fully-connected layers respectively. The matrix $\mathbf{W}$ is used to denote the parameters of the $K-1$ outputs and each column of $\mathbf{W}$ corresponds to the parameters of one output. The vector $\mathbf{x}_i$ denotes the output of the FC2 layer in MTNet for the given input face image $\mathbf{X}_i$. The sigmoid function is denoted as $\sigma(x)$, *i.e.*, $\sigma(x) = 1/(1+\exp(-x))$.

The term $\mathrm{cost}_k(y_i)$ in Eqn. (4) is used as an importance weight to rescale the training data $\mathbf{X}_i$ for the k-th output of the MTNet. As a result, the loss function is:

$$\operatorname*{argmin}_{\Omega} \sum_{i=1}^{N}\sum_{k=1}^{K-1}\Big\{-\mathrm{cost}_k(y_i)\Big(\mathbf{y}_i^{(k)}\log\sigma(\mathbf{W}^{(k)\,\mathrm{T}}\mathbf{x}_i) + (1-\mathbf{y}_i^{(k)})\log\big(1-\sigma(\mathbf{W}^{(k)\,\mathrm{T}}\mathbf{x}_i)\big)\Big)\Big\} + \sum_{j=1}^{M}\boldsymbol{\theta}_j^{\mathrm{T}}\boldsymbol{\theta}_j + \sum_{k=1}^{K-1}\mathbf{W}^{(k)\,\mathrm{T}}\mathbf{W}^{(k)} + \mathrm{tr}(\mathbf{W}_{\mathrm{FC1}}\mathbf{W}_{\mathrm{FC1}}^{\mathrm{T}}) + \mathrm{tr}(\mathbf{W}_{\mathrm{FC2}}\mathbf{W}_{\mathrm{FC2}}^{\mathrm{T}}), \tag{5}$$

where $\Omega = \{\Theta, \mathbf{W}_{\mathrm{FC1}}, \mathbf{W}_{\mathrm{FC2}}, \mathbf{W}\}$. The terms in the curly brackets correspond to the cost-sensitive multi-task loss, while the remaining terms represent the weight decay [18] of the MTNet's parameters. Weight decay is commonly used in deep learning to reduce overfitting. Even though the loss function in Eqn. (5) is a highly nonlinear function defined over the training data and the parameters of the MTNet, it can be efficiently solved in practice by the stochastic gradient descent algorithm [17].

In summary, at this cost-sensitive feature learning stage, age estimation is formulated as a ranking problem as it captures the correlations among age labels and learns more transferable aging features. An MTNet architecture (*c.f.* Figure 2) for age estimation is designed under this ranking based formulation. We then discuss the loss function of this MTNet and derive a cost-sensitive multi-task loss function (*c.f.* Eqn. (5)) for it. Based on the above knowledge, An MTNet Net^s is trained on the source population training data $\mathcal{X}^s$. We then create an MTNet Net^t for the target population by copying all parameters from Net^s. After this first stage, useful and transferable low-level aging features are extracted from the large sized source population data and then directly transferred to the target population. This direct feature transfer from the source to the target population exploits the ability of deep learning to capture hierarchical features independently of the training data, particularly from the lower layers [39, 31].

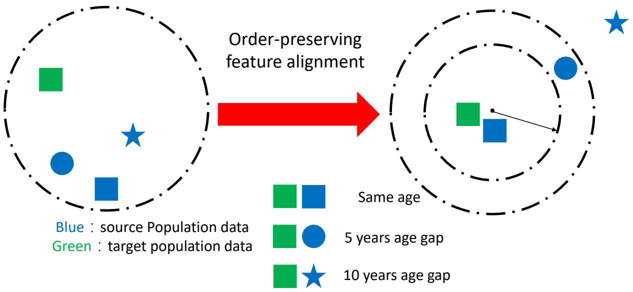
Order-preserving feature alignment
Same age
5 years age gap
10 years age gap
Blue : source Population data
Green : target population data

Figure 3. The key idea of the order-preserving feature alignment stage. There are four face images in this figure. The green one is from the target population, and the blue ones are from the source population. We can use these four images to construct three cross-population pairs. After this order-preserving feature alignment, the distance of the pair with the same age (the green and blue squares) becomes smaller, and those distances of the pairs with different ages become larger. Moreover, the pair with larger age gap (the green square and the blue star) has larger distance than that with smaller age gap (the green square and the blue circle). This figure is best viewed in color.

3.2. Order-preserving feature alignment stage

At this stage, we fine-tune both the source MTNet Net^s and the target MTNet Net^t obtained at stage one to transfer cross-population pair-wise information and perform incremental learning on the target population. The key idea is to align the high-level aging features from source and target populations to a *population-invariant space* which can capture the order characteristics of human ages. More specifically, in this population-invariant space, distances between face pairs with the same age are small, and those between pairs with different ages are large. Moreover, the pairs with larger age gap have larger distance than those with smaller age gap. In the following, we introduce our order-preserving feature alignment to achieve this goal.

Given a face image $\mathbf{X}_i^s$ with age label y_i^s from the source population training data $\mathcal{X}^s$, and another face image $\mathbf{X}_j^t$ with age label y_j^t from the target population training data $\mathcal{X}^t$, a cross-population pair $(\mathbf{X}_i^s, \mathbf{X}_j^t, y_i^s, y_j^t, l_{ij})$ is constructed, where l_{ij} is set to 1 if $y_i^s = y_j^t$ and -1 otherwise. The goal of our order-preserving feature alignment is to minimize the following objective function:

$$\sum_{i=1}^{N^s}\sum_{j=1}^{N^t}\{1 - l_{ij}(\eta - d(\hat{\mathbf{x}}_i^s, \hat{\mathbf{x}}_j^t)) \cdot \omega(y_i^s, y_j^t)\}. \tag{6}$$

Here $\hat{\mathbf{x}}_i^s$ and $\hat{\mathbf{x}}_j^t$ are the high-level aging features (*i.e.*, the vectorised feature maps of the Pool3 layer in Figure 2) extracted by Net^s and Net^t respectively. $d(\hat{\mathbf{x}}_i^s, \hat{\mathbf{x}}_j^t) = \|\hat{\mathbf{x}}_i^s - \hat{\mathbf{x}}_j^t\|_{\mathrm{F}}^2$ is the squared Euclidean distance between the high-level aging features, η is a pre-specified threshold parameter and is set to 2 experimentally. $\omega(y_i^s, y_j^t)$ denotes the

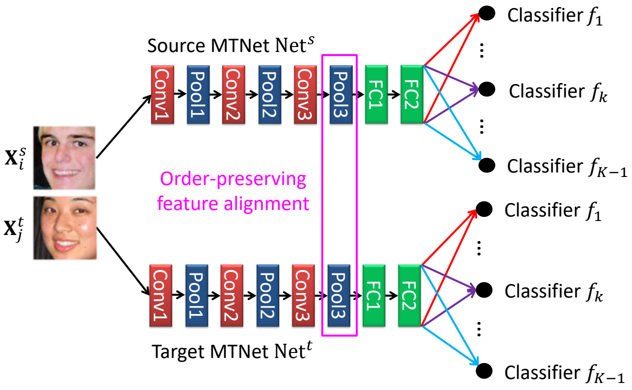

Figure 4. Order-preserving feature alignment stage. At this stage, both the source MTNet Net^s and the target MTNet Net^t are fine-tuned using cross-population face pairs, the relations of which are useful in bridging the population gap by transferring knowledge encoded in the source population to the target population which has only a small amount of training data. After this stage, the MTNet Net^t is ready to be deployed on the target population.

weighing function, which is computed as follows:

$$\omega(y_i^s, y_j^t) = \begin{cases} 1 - \exp(-\frac{|y_i^s - y_j^t|}{\tau}), & \text{if } y_i^s \neq y_j^t \\ 1, & \text{otherwise,} \end{cases} \quad (7)$$

where τ is a empirical pre-specified parameter which is set to 10 in the experiments. The rationales of Eqns. (6) and (7) are as follows (*c.f.* Figure 3).

- If a cross-population face image pair $\mathbf{X}_i^s$ and $\mathbf{X}_j^t$ have the same age, by the definition of Eqn. 6, the distance between them is expected to be as small as possible. Otherwise, the distance between them is expected to be as large as possible. As a result, the margin between pairs with the same age and pairs with different ages is maximized and discriminative aging feature is obtained in the aligned feature space.

- The larger the difference between y_i^s and y_j^t, the larger the weight $\omega(y_i^s, y_j^t)$ is assigned according to Eqn. (7). Then, by Eqn. (6), pairs with large age gaps are expected to have larger distances than pairs with smaller age gaps. As a result, the order characteristics of human ages are preserved in the aligned feature space.

Figure 4 shows the training details at this order-preserving feature alignment stage. During training, we use a mini-batch of cross-population pairs. For simplicity, only one cross-population pair $(\mathbf{X}_i^s, \mathbf{X}_j^t, y_i^s, y_j^t, l_{ij})$ is shown in Figure 4. We feed $\mathbf{X}_i^s$ and $\mathbf{X}_j^t$ to Net^s and Net^t respectively. The objective function of the order-preserving feature alignment in Eqn. (6) is used to align the aging features to a population-invariant space which captures the order characteristics of human ages. This alignment is useful in bridging the large population gap, so Net^t benefits from the large size source population data. Concurrently, the cost-sensitive multi-task loss function in Eqn. (5) is also used to fine-tune Net^t by using the target population training data $(\mathbf{X}_j^t, y_j^t)$.

In summary, in the cross-population age estimation setting, the aim is to optimize the MTNet Net^t for the target population. This is achieved during model training by learning a source MTNet Net^s with a cost-sensitive multi-task loss function on the large sized source population for transferring low-level aging features (Section 3.1), followed by an order-preserving feature alignment stage for transferring cross-population pairing knowledge and adapting the Net^t to data from the target population (Section 3.2). Now, Net^t can be used for age estimation for faces from the target population.

4. Experiments

In this section, the experimental settings are described in detail. Then, we conduct extensive experiments to validate the effectiveness of the proposed DCP age estimation model, with comparisons with the state-of-the-art and with a set of ablative studies.

4.1. Experimental settings

4.1.1 Datasets

There are many datasets for age estimation in the literature [21, 6, 4]. However, most of these datasets are relatively small. In order to obtain statistically meaningful results, we conduct experiments on two of the largest age estimation benchmark datasets, *i.e.*, the Morph II [29] and the WebFace [33] datasets in this work.

Morph II dataset: The Morph II dataset contains about $55,000$ face images of more than $13,000$ subjects with ages ranging from 16 to 77 years old. Morph II is a multi-ethnic dataset. It has about 77% Black faces and 19% White faces, while the remaining 4% are other races, *e.g.*, Hispanic, Indian, Asian. We followed the first cross-population age estimation study [13], and assembled a database of $21,060$ face images. More specifically, there are $7,960$ White Male (WM), $7,960$ Black Male (BM), $2,570$ White Female (WF), and $2,570$ Black Female (BF) face images in this assembled database. We treat WM/BM as the source population and WF/BF as the target population, which agrees with our cross-population age estimation setting in that the source population has more training data than the target population. The data of the target population (WF/BF) is randomly divided into two subsets with an equal size. One subset together with all of the source population data is used for training, while the other subset is used for testing.

WebFace dataset: The WebFace dataset contains $59,930$ face images with ages ranging from 1 to 80 years old. This dataset is also a multi-ethnic dataset, and most of the images are White or Yellow faces. In contrast with the

Morph II dataset which contains mug-shot face images, this dataset is compiled from face images captured in the wild. The images contain large pose and expression variations, which make this dataset much more challenging. In order to conduct cross-population age estimation experiments, we assembled a database of $34,000$ face images. Specifically, there are $14,000$ White Male (WM), $14,000$ White Female (WF), $3,000$ Yellow Male (YM), and $3,000$ Yellow Female (YF) face images in this assembled database. Similarly, we treat WM/WF as the source population and YM/YF as the target population. The data of the target population (YM/BF) is also randomly divided into two subsets with an equal size. One is used for training, while the other for testing.

4.1.2 Evaluation metric

To evaluate the performance of different age estimation algorithms, we use the popular Mean Absolute Error (MAE) as the evaluation metric. The MAE is calculated based on the average absolute error between the estimated age and the ground truth age, which is defined as follows,

$$\mathrm{MAE} = \frac{1}{N} \sum_{i=1}^{N} \left| \tilde{y}_i - y_i \right|, \tag{8}$$

where N is the number of testing face images, y_i is the ground-truth age of the i-th face image, and $\tilde{y}_i$ is the predicted age for it. Smaller MAE values mean better age estimation performance.

4.1.3 Parameter settings

The face images in both datasets are preprocessed following standard processing pipeline, *i.e*., the faces in the images are detected, aligned, and then cropped to 256×256 pixels. In all the following experiments, we use the Caffe [16] toolbox, which is a flexible deep learning framework to develop new models, and makes our work easy to reproduce. We train all the networks using stochastic gradient descent with momentum (0.9) and weight decay (5×10^{-4}). The dropout ratio is set to 0.5. The data augmentation strategy is similar to [17], *i.e*., randomly cropping of 227×227 pixels from the 256×256 input face image, then randomly flipping it before feeding it to the network. The initial learning rate is 10^{-3} which is divided by 10 when the training curve reaches a plateau. We found that all networks converge well under these settings, so we use the same hyper-parameters for different models to make fair comparisons.

4.1.4 Compared methods

We compare the DCP age estimation model with **two** state-of-the-art models and **five** deep baseline models. Since the cross-population age estimation is a relatively new problem, to the best of our knowledge, there are only two previous works which focus specifically on this problem: 1) Cross-population Discriminant Analysis (CpDA) [13], and 2) Joint Metric Learning (JML) [1].

Since the DCP model is the first deep learning based model for cross-population age estimation. In order to show its effectiveness, we design five deep baseline models for comparisons: 1) No Adaptation (NA). We train an MTNet using the source population data and directly deploy it for the target population testing data. This direct transfer scheme shows some success due to the generalization ability of deep models; 2) Direct Training (DT). We train an MTNet on the target population training data directly and then test it on the target population testing data; 3) United Populations (UP). We train an MTNet on the union of the source and target population training data. Compared with DT, more data are used for model training so that the performance may be improved; 4) Fine-tune based Transfer (FT). We first train an MTNet on the source population data, then fine-tune the fully-connected layers of it on the target population training data. This transfer learning strategy is widely used in the deep learning literature; and 5) Deep Joint Metric Learning (DJML). The aforementioned four deep baseline models are not specific to the cross-population age estimation problem. To get a stronger deep baseline mode, we reimplement the JML model [1] by incorporating the metric learning into the MTNet.

4.2. Comparison with the state-of-the-art models

We compare our DCP age estimation model with the state-of-the-art methods, *i.e*., CpDA and JML. For fair comparison, we conduct experiments on the Morph II dataset, since all of these three models are based on the same training and testing split protocol on this dataset. The results are shown in Table 1. Compared with CpDA and JML, the DCP age estimation model reduces the errors in each cross-population case significantly. For example, in the first cross-population case, the Black Male (BM) is used as the source population and the Black Female (BF) is used as the target population. The CpDA has a MAE of 7.73 years and the JML has a MAE of 5.56 years. Our DCP model reduces the MAE to 3.75 years which are 51.49% and 32.55% relative improvements respectively. Compared with the state-of-the-art methods which use handcrafted aging features and optimize each component independently, our DCP age estimation model can simultaneously learn aging features and an age estimator in an end-to-end framework and thus obtains superior performance.

4.3. Comparison with the deep baseline models

The DCP age estimation model is compared with the five deep baseline models. The cross-population age estima-

Table 1. Comparison with the state-of-the-art cross-population age estimation methods on the Morph II dataset.

Source	Target	CpDA [13]	JML [1]	DCP
BM	BF	7.73	5.56	**3.75**
	WF	8.73	5.57	**3.18**
WM	BF	7.67	6.40	**3.90**
	WF	6.70	5.00	**3.13**

Table 2. Comparison with the five deep baseline cross-population age estimation models on the Morph II dataset.

Source	Target	NA	DT	UP	FT	DJML	DCP
BM	BF	5.93	4.15	3.99	3.93	3.81	**3.75**
	WF	6.79	3.69	3.51	3.48	3.30	**3.18**
WM	BF	6.71	4.15	4.10	4.05	4.00	**3.90**
	WF	5.57	3.69	3.34	3.32	3.20	**3.13**

Table 3. Comparison with the five deep baseline cross-population age estimation models on the WebFace dataset.

Source	Target	NA	DT	UP	FT	DJML	DCP
WM	YM	6.78	5.45	5.24	4.75	4.73	**4.61**
	YF	10.06	5.69	5.32	4.91	4.80	**4.65**
WF	YM	9.24	5.45	5.15	4.82	4.72	**4.60**
	YF	7.41	5.69	4.55	4.49	4.40	**4.33**

Table 4. The age estimation results of the MTNet and MTNet (w/o cost-sensitive) on the Morph II dataset.

Source	MTNet	MTNet (w/o cost-sensitive)
BM	**3.37**	3.39
WM	**2.80**	2.84

Table 5. The age estimation results of the MTNet and MTNet (w/o cost-sensitive) on the WebFace dataset.

Source	MTNet	MTNet (w/o cost-sensitive)
WM	**6.64**	6.85
WF	**6.90**	7.17

tion results of these models on the Morph II and WebFace datasets are show in Tables 2 and 3 respectively.

The No Adaptation (NA) model has the largest MAE in each cross-population case. This is because different populations have different aging patterns, so the model trained on the source population can not perform well on the target population without any adaptations. From the results of NA, we can also see that the MAE when both race and gender are crossed is larger than the MAE when only race or only gender are crossed. For example, on the Morph II dataset, the cross-population case BM $\rightarrow$ BF has a MAE of 5.93 years, while BM $\rightarrow$ WF has a MAE of 6.79 years. This is because the aging patterns differences of the populations with different race and gender are larger than that of populations with either different race or different gender.

The Direct Training (DT) model performs better than NA because DT directly uses the target population training data for model training. From the results of DT, we can see that the MAE of BF is larger than the MAE of WF on the Morph II dataset. The main reason behind this is that it is easier to detect the facial appearance changes of White people than those of Black people. We can also see that the MAE of YF is larger than the MAE of YM on the WebFace dataset. This is because males and females have different face aging patterns. Many female faces appear younger than the male faces because of the use of makeup and accessories. This fact makes it more difficult to estimate the age of females [5, 36].

When the additional source population training data were utilised, the United Populations (UP) model has a better age estimation performance than the DT model. This supports the hypothesis that the source population data encodes useful knowledge which is beneficial for age estimation on the target population. From the results of UP, we can also observe that in most cross-population cases, the more similar the source population and the target population, the better the performance of the cross-population age estimation. For example, on the WebFace dataset, the cross-population case WF $\rightarrow$ YF has a MAE of 4.55 years which is better than the case WF $\rightarrow$ YM with a MAE of 5.15 years. The reason is that it is easier to transfer the knowledge encoded in the source population to the target population when they are similar.

We can see that the Fine-tune based Transfer (FT) model performs better than UP on both datasets and in each cross-population case. This demonstrates that FT is a better transfer strategy than UP for the cross-population problem. The reason is that UP trains on the union of the source and target population data directly. It makes the network difficult to learn since faces with the same age label may have different aging patterns if they come from different populations.

The DJML and our DCP model are specifically designed for the cross-population age estimation problem. We observe that they perform better than the previous four deep baseline models. This is because both DJML and our DCP model use the cross-population pair-wise information to align the aging features which is critical for the cross-population age estimation problem. We also see that our DCP age estimation model obtains the best performance on both datasets and in each cross-population case. This is because the DJML does not take the order characteristics of human ages into account, while the DCP model aligns the aging features of the source and target populations to a population-invariant space which captures the order characteristics of human ages. All of these experimental results and analyses demonstrate that the DCP model is effective for the cross-population age estimation problem.

4.4. Ablation Experiments

At last of this section, we conduct some ablative studies to further verify the effectiveness of each component

Table 6. The cross-population age estimation results of DCP and DCP$^-$ on the Morph II dataset in each cross-population case.

Source	Target	DCP	DCP$^-$
BM	BF	**3.75**	3.87
	WF	**3.18**	3.30
WM	BF	**3.90**	3.98
	WF	**3.13**	3.23

of the DCP age estimation model. More specifically, we show the effects of the cost-sensitive feature learning, the order-preserving feature alignment, and the target population training data size.

Effects of the cost-sensitive feature learning. The effects of the cost-sensitive learning defined in Equations 4 and 5 are evaluated. We conduct experiments on the source population data since it is relatively large in size. Specifically, we randomly divide the source population data into two subsets with an equal size. One is used for training, while the other for testing. Tables 4 and 5 show the experimental results on the Morph II and WebFace datasets respectively. It can be seen that the MTNet with cost-sensitive learning obtains better performance on both datasets. These experimental results demonstrate that incorporating the inherent cost sensitivity of age estimation into model training improves the age estimation performance.

Effects of the order-preserving feature alignment. The effects of the order-preserving feature alignment defined in Equations 6 and 7 are evaluated. In order to show the effectiveness of the alignment, we make a comparison with DCP$^-$ which does not take the order characteristics of human ages into consideration. More specifically, in DCP$^-$, the weighing function defined in Eqn. 7 always equals to 1. The experimental results of these two models on the Morph II and WebFace datasets are shown in Tables 6 and 7 respectively. The DCP outperforms DCP$^-$ on both datasets and in every cross-population case. This is because DCP$^-$ separates pairs with different ages equally without taking into consideration the difference in their ages. For example, two pairs of face images with the ages $(20, 50)$ and $(20, 21)$ are pushed apart equally which is unsatisfactory since faces with neighbouring ages are generally more similar in appearance than faces with widely separated ages. In contrast, in the order-preserving feature alignment of our DCP model, the pair with a larger age gap is expected to have larger distance than that with smaller age gap. As a result, the order characteristics of human ages is preserved and thus better performance is obtained.

Effects of the target population training data size. The cross-population age estimation performance of the DCP model is evaluated for a range of target population training data sizes. The purpose of this experiment is to answer if a smaller number of face images in the target population can be sufficient for learning. To this end, we reduce the number of target population training face images to $\{90\%, 80\%, 70\%, 60\%, 50\%, 40\%, 30\%, 20\%, 10\%\}$ of the full training set. The results are shown in Figure 5. As expected, the performance of our DCP age estimation model degrades when target population training data are removed. But, we also observe that a small amount of the target population training data is sufficient to learn our DCP model with a good performance. For example, about 30% of the target population training data is enough to obtain a MAE which is within one year difference from the 100% training data. This is very useful in practice, because only a small amount of target training data is required to obtain satisfactory age estimation performance.

Table 7. The cross-population age estimation results of DCP and DCP$^-$ on the WebFace dataset in each cross-population case.

Source	Target	DCP	DCP$^-$
WM	YM	**4.61**	4.72
	YF	**4.65**	4.75
WF	YM	**4.60**	4.74
	YF	**4.33**	4.43

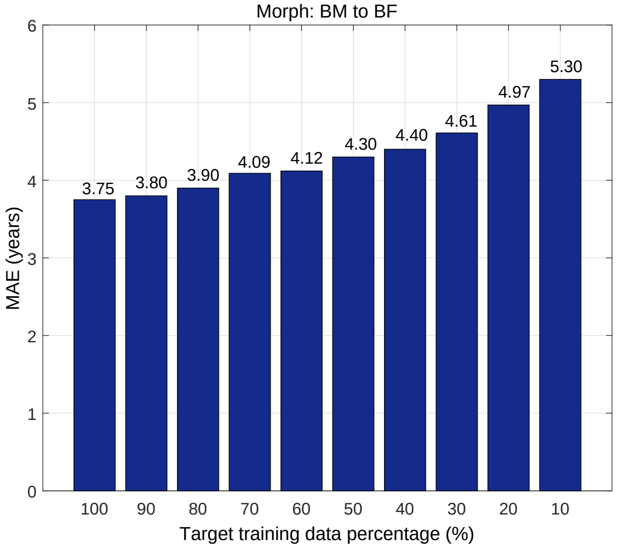

Figure 5. The cross-population age estimation results w.r.t. the percentage of the training data in the target population.

5. Conclusions and Future Work

In this paper, we have proposed a DCP model for the challenging cross-population age estimation problem. The model includes two training stages. In the first stage, age estimation is formulated as a ranking problem and a novel cost-sensitive multi-task loss function is designed, to learn transferable low-level aging features on the source population. In the second stage, a novel order-preserving feature alignment procedure is designed to align the high-level aging features, and simultaneously include the target population data in the training process. After this two-stage training, the DCP model effectively transfers the knowledge encoded in the source population to the target population. The DCP model has been evaluated on the two of the largest age estimation datasets. The experimental results show that the DCP model is more accurate than two state-of-the-art methods and five deep baseline models. In the future work, we plan to study the cross-population age estimation problem when there are no labeled data in the target population.

References

[1] B. Bhattarai, G. Sharma, A. Lechervy, and F. Jurie. A joint learning approach for cross domain age estimation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1901–1905, 2016. 1, 6, 7

[2] K.-Y. Chang and C.-S. Chen. A learning framework for age rank estimation based on face images with scattering transform. *IEEE Transactions on Image Processing*, 24(3):785–798, 2015. 1, 3

[3] Q. Dong, S. Gong, and X. Zhu. Multi-task curriculum transfer deep learning of clothing attributes. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, pages 1–8, 2017. 1

[4] E. Eidinger, R. Enbar, and T. Hassner. Age and gender estimation of unfiltered faces. *IEEE Transactions on Information Forensics and Security*, 9(12):2170–2179, 2014. 5

[5] Y. Fu, G. Guo, and T. S. Huang. Age synthesis and estimation via faces: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(11):1955–1976, 2010. 1, 7

[6] Y. Fu and T. S. Huang. Human age estimation with regression on discriminative aging manifold. *IEEE Transactions on Multimedia*, 10(4):578–584, 2007. 5

[7] X. Geng, Z.-H. Zhou, and K. Smith-Miles. Automatic age estimation based on facial aging patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(12):2234–2240, 2007. 1

[8] X. Geng, Z.-H. Zhou, Y. Zhang, G. Li, and H. Dai. Learning from facial aging patterns for automatic age estimation. In *Proceedings of the ACM International Conference on Multimedia*, pages 307–316, 2006. 1

[9] 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. 1

[10] G. Guo and G. Mu. Human age estimation: What is the influence across race and gender? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 71–78, 2010. 1

[11] G. Guo, G. Mu, Y. Fu, C. Dyer, and T. Huang. A study on automatic age estimation using a large database. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1986–1991, 2009. 1

[12] G. Guo, G. Mu, Y. Fu, and T. Huang. Human age estimation using bio-inspired features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 112–119, 2009. 1

[13] G. Guo and C. Zhang. A study on cross-population age estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4257–4263, 2014. 5, 6, 7

[14] H. Han, C. Otto, X. Liu, and A. K. Jain. Demographic estimation from face images: Human vs. machine performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(6):1148–1161, 2015. 1

[15] J. Hu, J. Lu, and Y.-P. Tan. Deep transfer metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 325–333, 2015. 1

[16] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the ACM International Conference on Multimedia*, pages 675–678, 2014. 6

[17] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 1097–1105, 2012. 4, 6

[18] A. Krogh and J. A. Hertz. A simple weight decay can improve generalization. In *In Proceedings of Advances in Neural Information Processing Systems*, pages 950–957, 1992. 4

[19] Z. Kuang, C. Huang, and W. Zhang. Deeply learned rich coding for cross-dataset facial age estimation. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 96–101, 2015. 3

[20] Y. H. Kwon and N. da Vitoria Lobo. Age classification from facial images. *Computer Vision and Image Understanding*, 74(1):1–21, 1999. 1

[21] A. Lanitis, C.Taylor, and T. Cootes. Toward automatic simulation of aging effects on face images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(4):442–455, 2002. 1, 5

[22] A. Lanitis, C. Draganova, and C. Christodoulou. Comparing different classifiers for automatic age estimation. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 34(1):621–628, 2004. 1

[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. 1, 3

[24] K. Li, J. Xing, W. Hu, and S. J. Maybank. D2C: Deep cumulatively and comparatively learning for human age estimation. *Pattern Recognition*, 66(C):95–105, 2017. 1, 3

[25] L. Li and H.-T. Lin. Ordinal regression by extended binary classification. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 865–872, 2007. 3

[26] J. Lu, V. E. Liong, and J. Zhou. Cost-sensitive local binary feature learning for facial age estimation. *IEEE Transactions on Image Processing*, 24(12):5356–5368, 2015. 1

[27] Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua. Ordinal regression with multiple output CNN for age estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4920–4928, 2016. 3

[28] J. K. Pontes, A. S. Britto, C. Fookes, and A. L. Koerich. A flexible hierarchical approach for facial age estimation based on multiple features. *Pattern Recognition*, 54(C):34–51, 2016. 1

[29] K. Ricanek and T. Tesafaye. Morph: A longitudinal image database of normal adult age-progression. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, pages 341–345, 2006. 5

[30] C. Shan, F. Porikli, T. Xiang, and S. Gong, editors. *Video Analytics for Business Intelligence*. Studies in Computational Intelligence. Springer, 2012. 1

[31] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. CNN features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 806–813, 2014. 2, 4

[32] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the International Conference on Learning Representations*, pages 1–14, 2014. 3

[33] Z. Song. *Visual Image Recognition System with Object-Level Image Representation*. PhD thesis, National University of Singapore, 2012. 5

[34] Z. Song, B. Ni, D. Guo, T. Sim, and S. Yan. Learning universal multi-view age estimator using video context. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 241–248, 2011. 1

[35] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015. 3

[36] J. Xing, K. Li, W. Hu, C. Yuan, and H. Ling. Diagnosing deep learning models for high accuracy age estimation from a single image. *Pattern Recognition*, 66(C):106–116, 2017. 1, 3, 7

[37] S. Yan, X. Zhou, M. Liu, M. Hasegawa-Johnson, and T. S. Huang. Regression from patch-kernel. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008. 1

[38] H.-F. Yang, B.-Y. Lin, K.-Y. Chang, and C.-S. Chen. Automatic age estimation from face images via deep ranking. In *Proceedings of the British Machine Vision Conference*, pages 1872–1886, 2015. 3

[39] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? In *Proceedings of the Advances in Neural Information Processing Systems*, pages 3320–3328, 2014. 2, 4

[40] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *Proceedings of the European Conference on Computer Vision*, pages 818–833, 2014. 3