Cross-Age Face Verification by Coordinating with Cross-Face Age Verification

Liang Du Haibin Ling*
Department of Computer and Information Sciences, Temple University, Philadelphia, USA
{liang.du, hbling}@temple.edu

Abstract

In this paper we present a novel framework for cross-age face verification *(FV) by seeking help from its "competitor" named* cross-face age verification *(AV), i.e., deciding whether two face photos are taken at similar ages. While FV and AV share some common features, FV pursues* age insensitivity *and AV seeks* age sensitivity. *Such correlation suggests that AV may be used to guide feature selection in FV, i.e., by reducing the chance of choosing age sensitive features. Driven by this intuition, we propose to learn a solution for cross-age face verification by coordinating with a solution for age verification. Specifically, a joint additive model is devised to simultaneously handling both tasks, while encoding feature coordination by a competition regularization term. Then, an alternating greedy coordinate descent (AGCD) algorithm is developed to solve this joint model. As shown in our experiments, the algorithm effectively balances feature sharing and feature exclusion between the two tasks; and, for face verification, the algorithm effectively removes distracting features used in age verification. To evaluate the proposed algorithm, we conduct cross-age face verification experiments using two benchmark cross-age face datasets, FG-Net and MORPH. In all experiments, our algorithm achieves very promising results and outperforms all previously tested solutions.*

1. Introduction

Facial image analysis has a wide range of applications such as visual surveillance, human computer interactions, and biometric verification. As an important factor in facial image analysis, human age is gaining increasing attention especially in two topics: cross-age face identity analysis and age inference. These two topics can be viewed as competing with each other: one seeks age *insensitivity* while the other age *sensitivity*. Previously the tasks in these two topics are investigated independently, while in this paper we explore the benefit of the coordination between them.

*Correspondence author.

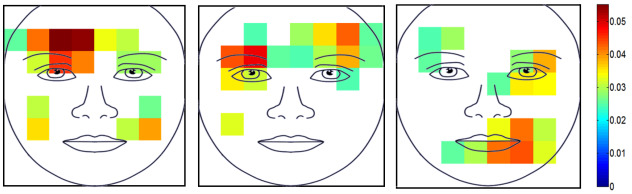


Figure 1. Feature importance map (spatially projected): (a) Baseline face verification (FV) without coordination; (b) age verification (AV) coordinating with FV; (c) FV coordinating with AV. The baseline FV uses a large amount of features around the forehead region. While in our joint learning solution, such age-sensitive features are effectively inhibited (c) by coordinating with AV (b).

Intuitively, knowing *a priori* that two tasks have conflicting goals can help inhibit irrelevant features for each task and hence improve the performance. For example, some critical features for age estimation, *e.g.*, wrinkles on forehead, may be irrelevant for cross-age face verification and can be safely excluded [19]. Inhibition of such features may reduce the risk of over-fitting, an annoying issue for cross-age face verification that involves many sources of uncertainties but often has limited training data. This intuition suggests us to improve cross-age face verification algorithms by incorporating information from age inference, *i.e.*, identifying and avoiding age sensitive features in a principled way. The intuition is validated in our experimental analysis as shown in Fig. 1.

While the above observation is interesting, it is worth emphasizing that the exclusion of features should not be performed at a macro level[1], *e.g.*, excluding a spatial face area, but at a fine-grained level, *i.e.*, some feature dimensions of a specific type of features. Using the wrinkles on forehead as an illustrative example, for FV, it is desirable to exclude features which are discriminative for wrinkles rather than exclude all features from the whole forehead. Moreover, aside from the confliction in some age sensitive features, the two tasks may still share some features. For ex-

[1]By macro level feature sharing, we mean features from the same specific type or spatial area. For example, using different feature responses from the same pixel location is not counted as "sharing".

ample, some appearance features around eyes may encode information for both FV and AV. Consequently, we need a smart strategy to coordinate the two tasks that does not forbid feature sharing. Such feature sharing can be observed in Fig. 1 as well.

Based on the above discussion, we propose a novel framework to learn cross-age face verification solutions by coordinating with cross-face age verification (*i.e.*, deciding whether two facial photos were captured at similar ages). Sharing the same feature pool, the two tasks are modeled together in a joint loss framework, with feature interaction encouraged via an orthogonal regularization over feature importance vectors. Then, an alternating greedy coordinate descent learning algorithm (AGCD) is derived to estimate the model. The algorithm effectively excludes distracting features in a fine-grained level for improving face verification. In other words, the proposed algorithm does not forbid feature sharing between conflicting tasks at the macro level; it instead selectively inhibits distracting features while preserving discriminative ones, as analyzed in Sec. 5.3.

For evaluation, the proposed algorithm is applied to two widely tested face-aging benchmark datasets: FG-Net [6] and MORPH [29]. On both datasets, our algorithm achieves very promising performances and outperforms all previously reported results. These experiments, together with detailed experimental analysis, show clearly the benefit of coordinating conflicting tasks for improving visual recognition. In summary, we make three main contributions:

- To the best of our knowledge, our study is the first one that treats age-sensitive information as a blessing rather than a curse for cross-age face verification.
- A novel framework is proposed to harness the task conflict in a principled way, and a new algorithm is developed in this framework. The algorithm is general and can be extended to other scenarios involving similar task competition constraints.
- Extensive experiments on benchmark datasets are conducted and new results are registered.

In the rest of the paper, we first review related studies in Sec. 2. Then, in Sec. 3 we present the proposed joint modeling framework along with the developed new algorithm. In Sec. 4 we apply the algorithm to cross-age face verification. After that, we describe the experimental results in Sec. 5, followed by conclusion in Sec. 6.

2. Related Work

2.1. Cross-age face recognition

Face recognition in general is one of the most intensively researched computer vision topics [32, 43]. By contrast, recognition of facial images taken at different ages, despite its wide range of potential applications, has been under-explored. Recently, advances in face-aging datasets (e.g., [4, 6, 29]) largely boost the study along this line. In the following we roughly divide previous work on cross-age face recognition into two classes: generative and discriminative. We also leave out a relevant topic named age estimation, for which we refer the readers to a survey [9].

In generative solutions, the aging process is typically simulated and applied to transform a facial image of one age to the target age to reduce the aging effect for face recognition. In [27] a craniofacial growth model is used to predict facial appearance across years and then perform face recognition across-age progression for individuals under age 18. 3D aging model is used in [26] to compensate for the age variations to improve the face recognition. In [34], a compositional and dynamical model is proposed for face aging using the And-Or graph model. In [36], an approach is derived to synthesize face representation at the target age in the feature space before performing face recognition. The main challenges faced by generative solutions include the difficulties in modeling the complicated aging progress, and in estimating the target age accurately.

Discriminative methods usually focus on image descriptors or classifiers which are robust against age progression. An age difference classifier [28] in the Bayesian framework is used for verifying passport photos taken at different ages, given that the age difference of a pair is known a priori. Later on, gradient orientation pyramids (GOP) [20] is proposed for the similar task. In [12], the relationship between recognition accuracy and age intervals is investigated for soft biometric traits. In [18] popular local descriptors and multi-feature discriminative analysis (MFDA) are combined for cross-age face classification. A similar idea using multiview discriminative learning is explored in [33]. Subspace factor analysis is used to achieve age invariant for face recognition in [10]. Recently, Chen *et al*. proposed a coding framework for cross age face recognition by leveraging a large-scale image dataset [4]. A recent evaluation of several local descriptors is conducted in [3] on age-invariant face recognition. The results show that no single descriptor is versatile in discriminating faces with different age gaps.

Being a discriminative method, our algorithm is different from previous ones by seeking guidance from age inference. This is the first time the two tasks are jointly modeled to the best of our knowledge, and its effectiveness is empirically validated in comparison with previous solutions.

2.2. Related Work on Joint Task Learning

The benefits of learning with auxiliary tasks have been well recognized in the machine learning community [11, 25]. The proposed learning framework can be viewed as a multi-task learning one in that it models jointly two tasks: face verification and age verification. However, unlike most multiple task or transfer learning methods that explore similarities (*i.e.*, sharing of feature, instances or weak learn-

Figure 2. Example images (after alignment) used in our experiments. Each pair shows two photos of the same person captured at different ages (labeled under the photos).

ers), our method explores the competing relationship that discourages feature sharing by the nature of the tasks.

Recently, between-task competing or exclusion relationships have been drawing research attention in and exploited to regularize learning algorithms [5, 16, 21, 31, 44, 45]. The approaches in [16, 44, 45] rely on the top-down taxonomy and exploiting the feature competition relationships for subclass object classification. [31] proposed a linear model regularized by orthogonal relationships between model coefficients, for the tasks of expression recognition. In [5], competing relationships between tasks are instantiated during the boosting process when selecting weak learners for the tasks of expression recognition and writer identification. In [21], expression specific features are selected using exclusion constraints through a sparse support vector machine framework.

Our study are different than these studies in two main aspects. First, we emphasize on coordinating two conflicting tasks by suppressing distracting features identified by the conflicting relationship, through an alternating greedy coordinate descent algorithm. Second, we work on cross-age face verification to which previous studies have not been touched. Our study shares philosophies of joint boost algorithms used in vision such as [35]. Other examples of joint task learning in vision can be found in [41] for visual recognition and in [14] for joint age effect analysis.

3. Joint Task Modeling for Task Coordination

We start by formulating the problems of (cross-age) *face verification* and (cross-face) *age verification*. Given a pair of input facial images, (I_1, I_2), we first extract a d-dimensional feature vector from them, denoted by $\mathbf{x} \in \mathbb{R}^d$. Then, face verification and age verification are denoted respectively by functions $f : \mathbb{R}^d \to \{-1, +1\}$ and $f_a : \mathbb{R}^d \to \{-1, +1\}$, in which $+1$ indicates same identity for face verification or different age for age verification. We call face verification as the *target task* and age verification the *auxiliary task*.

In this section we first introduce the greedy coordinate descent strategy as a baseline solution. After that we model the target task and the auxiliary task jointly via a regularized joint loss, and then derive our solution via greedy coordinate descent. The details on applying the algorithm for face verification will be described in the next section.

3.1. Learning by Greedy Coordinate Descent

We first review the greedy coordinate descent strategy used for function approximation [8]. Let $\mathcal{D} = \{(\mathbf{x}_i, y_i) \in \mathbb{R}^d \times \{-1, +1\} : i = 1, \ldots, N\}$ be the training set of size N for the target task (face verification). Denote the pool of base hypothesis[2] as $\mathcal{H} = \{h_j : \mathbb{R}^d \to \{+1, -1\}, j = 1, \ldots, N_h\}$ of size N_h. The target classifier is modeled as a weighted combination of $\mathcal{H}$ as[3] $f(\mathbf{x}; \mathbf{p}) = \sum_{j=1}^{N_h} p_j h_j(\mathbf{x})$, where $\mathbf{p} = (p_1, p_2, \ldots, p_{N_h})^\top$ are the model coefficients. By enforcing that $h \in \mathcal{H} \Rightarrow -h \in \mathcal{H}$, we have $p_j \geq 0$, $j = 1, \ldots, N_h$.

The goal is to learn the model coefficients $\mathbf{p}$ to minimize a loss function of the form:

$$\mathcal{L}(\mathcal{D}, \mathbf{p}) = \sum_{i=1}^{N} \ell(y_i, f(\mathbf{x}_i; \mathbf{p})), \tag{1}$$

where $\ell(\cdot, \cdot)$ is the loss function for an individual sample.

By viewing each hypothesis as a coordinate and updating $\mathbf{p}$ along one coordinate in each iteration, various boosting algorithms can be derived by using iterative *greedy coordinate descent* (GCD) procedures to minimize (1) [8]. More specifically, in each iteration, the coordinate (base hypothesis) with maximum gradient is found via greedy coordinate descent, then the coefficient in $\mathbf{p}$ corresponding to the coordinate is updated with a stepsize, which can be either predefined or dynamically adjusted.

It has been shown that the popular AdaBoost algorithm [7] is equivalent to GCD procedure with an exponential loss $\ell(y, f(\mathbf{x}; \mathbf{p})) = \exp(-yf(\mathbf{x}; \mathbf{p}))$. In our implementation, we use the same exponential loss function and therefore our baseline GCD algorithm is essentially a boosting algorithm.

3.2. Link Base Hypotheses with Facial Features

Our goal is to explore the coordination between features instead of base hypotheses. For this reason, the

[2] Hereafter we treat base hypothesis and weak classifier synonymously.

[3] We use the soft version here for the learning procedure. For classification, the sign of the function is used, i.e., $f(\mathbf{x}; \mathbf{p}) = \text{sign}(\sum_{j=1}^{N_h} p_j h_j(\mathbf{x}))$.

weights in $\mathbf{p}$ need to be connected to the importance of features as in [40]. To make this association more clear, we use decision stumps for the base hypotheses. Specifically, the j-th ($1 \leq j \leq N_h$) weak hypothesis has the form $h_j(\mathbf{x}) = h_j(\mathbf{x}; s(j), \tau_j)$, indicating that h_j acts on the $s(j)$-th feature with stump parameter τ_j. Both $s(j)$ and τ_j are learnt from training samples. Using $s(j)$, the mapping from base hypotheses to features can be characterized by a binary *hypothesis-feature mapping matrix*, denoted as $M = (m_{kj}) \in \{0,1\}^{d \times N_h}$, such that $m_{kj} = \delta(s(j) - k) = 1$ if and only if $s(j) = k$. M is then used to map hypothesis weight vector $\mathbf{p}$ to a *feature importance vector*, denoted as $\boldsymbol{\phi}$, i.e., $\boldsymbol{\phi} = M\mathbf{p}$.

Intuitively, the k-th element in $\boldsymbol{\phi}$ is a weighted count of base hypotheses that act on the k-th feature. Since $\mathbf{p}$ and M are both nonnegative, so is $\boldsymbol{\phi}$. As demonstrated in the following subsection, the above mapping allows us to regularize the greedy coordinate descending process by taking into account feature coordination between tasks.

3.3. Regularized Joint Loss

Now we describe the joint modeling for the target task (face verification) and the auxiliary task (age verification). For the auxiliary task, we denote its training set by $\mathcal{A} = \{(\mathbf{z}_i, l_i) \in \mathbb{R}^d \times \{-1, +1\} : i = 1, \ldots, N_a\}$ of size N_a. Sharing the same pool of base hypothesis $\mathcal{H}$ with the target task, the auxiliary classifier has the form $f_a(\mathbf{x}; \mathbf{a}) = \sum_{j=1}^{N_h} a_j h_j(\mathbf{x})$, where $\mathbf{a} = (a_1, a_2, \ldots, a_{N_h})^\top$ are the nonnegative model coefficients.

Since both the target and the auxiliary tasks take similar forms and share the same hypothesis pool, we can define their joint loss as

$$\mathcal{J}_0(\mathcal{D}, \mathcal{A}, \mathbf{p}, \mathbf{a}) = \mathcal{L}(\mathcal{D}, \mathbf{p}) + \mathcal{L}(\mathcal{A}, \mathbf{a}) \,. \tag{2}$$

The above loss function allows the two tasks to share features extracted from the same pool. That said, we are also interested in modeling the conflict between the target task and the auxiliary task in the feature-level. We use M_p and M_a to denote the hypothesis-feature mapping matrices for the target and auxiliary tasks respectively; and similarly use $\boldsymbol{\phi}_p$ and $\boldsymbol{\phi}_a$ for their feature importance vectors, as described in the previous subsection. Then an orthogonal regularizer can be introduced over $\boldsymbol{\phi}_p$ and $\boldsymbol{\phi}_a$ to encode feature conflict between the two tasks. Then, the regularized joint loss is:

$$\begin{aligned} \mathcal{J}(\mathcal{D}, \mathcal{A}, \mathbf{p}, \mathbf{a}) &= \mathcal{J}_0(\mathcal{D}, \mathcal{A}, \mathbf{p}, \mathbf{a}) + \lambda \boldsymbol{\phi}_c^\top \boldsymbol{\phi}_a \\ &= \sum_{i=1}^{N} \exp(-y_i f(\mathbf{x}_i; \mathbf{p})) \\ &\quad + \sum_{i=1}^{N_a} \exp(-l_i f_a(\mathbf{z}_i; \mathbf{a})) + \lambda \mathbf{p}^\top M_c^\top M_a \mathbf{a} \,, \end{aligned} \tag{3}$$

where λ is the regularization parameter. Since both $\boldsymbol{\phi}_p$ and $\boldsymbol{\phi}_a$ are nonnegative, $\boldsymbol{\phi}_p^\top \boldsymbol{\phi}_a \geq 0$. This property is crucial for the success of orthogonal regularization in feature coordination. Without the nonnegativity, orthogonal regularization by itself can not lead to feature competition, since negative coefficients in a linear model may also contribute to feature importance.

3.4. Algorithm

We use the alternating greedy coordinate descent strategy to derive an algorithm for minimizing the regularized joint loss in (3). In each iteration, we first fix $\mathbf{a}$ to update $\mathbf{p}$, and then we fix $\mathbf{p}$ to update $\mathbf{a}$. The update of $\mathbf{p}$ or $\mathbf{a}$ is determined by finding the coordinate along which the loss decreases fastest.

When fixing $\mathbf{a}$, the partial derivative of $\mathcal{J}(\mathbf{p}, \mathbf{a})$[4] w.r.t. p_j is calculated by

$$\begin{aligned} -\frac{\partial \mathcal{J}(\mathbf{p}, \mathbf{a})}{\partial p_j} &= \sum_{i=1}^{N} \Big(y_i h_j(\mathbf{x}_i) \exp\big(- y_i \sum_{k=1}^{N_h} p_k h_k(\mathbf{x}_i) \big) \Big) \\ &\quad - \lambda \frac{\partial (\mathbf{p}^\top M_p^\top M_a \mathbf{a})}{\partial p_j} \,. \end{aligned} \tag{4}$$

Given the binary property of M_p, we have

$$\frac{\partial (\mathbf{p}^\top M_p^\top M_a \mathbf{a})}{\partial p_j} = (M_p(:, j))^\top \boldsymbol{\phi}_a = \boldsymbol{\phi}_a(s_p(j)) \,, \tag{5}$$

where $M_p(:, j)$ is the j-th column of M_p, $\boldsymbol{\phi}_a(s_p(j))$ the $s_p(j)$-th element of $\boldsymbol{\phi}_a$, and $s_p(j)$ the feature selection of the j-th hypothesis in f. Finally, we have

$$\begin{aligned} -\frac{\partial \mathcal{J}(\mathbf{p}, \mathbf{a})}{\partial p_j} &= \sum_{i=1}^{N} \Big(y_i h_j(\mathbf{x}_i) \exp\big(- y_i \sum_{k=1}^{N_h} p_k h_k(\mathbf{x}_i) \big) \Big) \\ &\quad - \lambda \boldsymbol{\phi}_a(s_p(j)) \,. \end{aligned} \tag{6}$$

Following the GCD strategy, we choose p_j that maximizes this term. This is similar to the base hypothesis selection under the current weight distribution in the classical AdaBoost algorithm, except that in our case there is an additional term $-\lambda \boldsymbol{\phi}_a(s_p(j))$. Therefore, the base hypothesis selection is not only to fit the current distribution of data, but also to discourage choosing features favored by the auxiliary task.

A similar alternating step is conducted for the auxiliary task learning, taking the partial derivative of $\mathcal{J}(\mathbf{p}, \mathbf{a})$ w.r.t. a_j as

$$\begin{aligned} -\frac{\partial \mathcal{J}(\mathbf{p}, \mathbf{a})}{\partial a_j} &= \sum_{i=1}^{N_a} \Big(l_i h_j(\mathbf{z}_i) \exp\big(- l_i \sum_{k=1}^{N_h} a_k h_k(\mathbf{z}_i) \big) \Big) \\ &\quad - \lambda \boldsymbol{\phi}_p(s_a(j)) \,, \end{aligned} \tag{7}$$

[4] For notation conciseness, we ignore $\mathcal{D}$ and $\mathcal{A}$ hereafter.

Algorithm 1 Alternating Greedy Coordinate Descent for Coordinative Task Learning(AGCD)

Require: Training sets: $\mathcal{D}$ for the target task and $\mathcal{A}$ for the auxiliary task

1: $\mathbf{p} \leftarrow \mathbf{0}, \mathbf{a} \leftarrow \mathbf{0}$
2: **for** $t = 1$ to T **do**
3: Fix $\mathbf{a}$ and find the "best" coordinate h_{j_t} according to (6):
$$j_t \leftarrow \arg \max_{j:1\leq j\leq N_h} -\frac{\partial \mathcal{J}(\mathbf{p}, \mathbf{a})}{\partial p_j}$$
4: Update the coefficient:
$$p_{j_t} \leftarrow p_{j_t} + \epsilon$$
5: Fix $\mathbf{p}$ and find the "best" coordinate h_{k_t} according to (7):
$$k_t \leftarrow \arg \max_{k:1\leq k\leq N_h} -\frac{\partial \mathcal{J}(\mathbf{p}, \mathbf{a})}{\partial a_k}$$
6: Update the coefficient:
$$a_{k_t} \leftarrow a_{k_t} + \epsilon$$
7: **end for**
8: **return** $f(\mathbf{x}; \mathbf{p}) = \sum_{j=1}^{N_h} p_j h_j(\mathbf{x})$

where $s_a(j)$ is the feature selection of the j-th hypothesis in f_a.

By using the same exponential loss function as AdaBoost, the proposed alternating greed coordinate descent algorithm over the regularized joint loss (3) shares a similar form with the classical AdaBoost algorithm. The AGCD algorithm is summarized in Algorithm 1. Compared with the GCD baseline, the only additional overhead is incurred by (5) which can be calculated very efficiently by updating $\boldsymbol{\phi}_a$ and $\boldsymbol{\phi}_p$ in each iteration.

The stepsize ϵ can be either fixed (typically small values) or greedy selected in each iteration. In addition, different ϵ's can be used for the target and auxiliary tasks. In our implementation, we fix ϵ as 0.1 in all experiments by following the empirical and theoretical conclusion that small stepsizes often help in boosting [8, 42].

4. Cross Age Face Verification using AGCD

The AGCD algorithm provides a general framework to improve the learning of a target task by coordinating with an auxiliary one. Though the algorithm by itself is symmetric for the target and auxiliary tasks, in practice, we often focus on the target task. In particular, in this paper we focus on cross-age face verification and design an auxiliary task that seeks age sensitive information.

4.1. Auxiliary Task Design

For cross-age face verification, a natural choice of a conflicting auxiliary task is age estimation, which explicitly pursues age sensitive information. Unfortunately, the task does not align directly with face verification where the input is an image pair, neither does it share the same type of output.

To address this issue, we instead choose *age verification*, which also takes an image pair as input and generates binary output (similar age or not). Age verification is also known as *age gap classification*, *i.e.*, an age difference is treated as positive if it is larger than a predefined threshold, and otherwise negative.

An important issue that needs special care is the covariation between the target task labels and auxiliary task labels in the training samples. Ideally, correlation between the labels of the two tasks should be reduced as much as possible. In the context of our study, the distributions of ages and identities should be kept uncorrelated. More specifically, a specific age label should not be associated with any individuals, regardless of gender, race and so on. In practice, however, due to the limitation of available data, there is inevitable covariation between the labels of the two tasks. This issue, if not handled properly, could neutralize our assumption and hurt the performance of our algorithm.

We attack this problem by intentionally remove such covariation as much as possible in the training dataset. With this de-correlation, the training set for age verification (the auxiliary task) contains many extra-personal image pairs. In other words, the age verification task is indeed *cross-face*. Details of the auxiliary tasks for each dataset are presented in the experimental section.

4.2. Feature Pool for Cross Age Face Verification

We follow a commonly used procedure to prepare a feature pool for face verification. For an input facial image, it is first aligned according to the eye positions and then cropped and converted to a greyscale image of size 64×64. For images in FG-Net, the manually labeled facial landmarks, including eye positions, are available. For images in MORPH, we use the tool in [1] for eye detection[5].

After the above preprocessing, we construct a feature pool by concatenating multiple features from multiple overlapping facial blocks. Four types of features used in our study include the *scale invariant feature transform* (SIFT) [23], *local binary pattern* (LBP) [2], *gradient orientation pyramid* (GOP) [20], and biological inspired features (BIF) [13, 15, 30]; all of them have been used recently for face analysis. For SIFT and LBP, following the densely sampling strategy in [18, 33], each face is divided into 49

[5]Code available: http://ibug.doc.ic.ac.uk/resources/drmf-matlab-code-cvpr-2013/

overlapping blocks of size 16×16, with an stepsize of 8 pixels, and the features are then extracted from each block. This results in $18,816$ $(= (128 + 256) \times 49)$ features for each image. For an image pair, the difference between their concatenated SIFT and LBP features is used as its representation. The GOP features collected the gradient orientations at multiple scales. For an image pair, its representation contains the cosines of the differences between gradient orientations at all pixels over multiple scales, leading to a feature dimension of $5,376$ $(= 64 \times 64 + 32 \times 32 + 16 \times 16)$. For BIF, we use 8 orientations and 12 scales as suggested in [13], with the "MAX" operator for pooling [30]. PCA is applied to reduce the dimension of BIF feature to $1,000$.

Finally, there are in total $25,192$ $(= 18,816 + 5,376 + 1,000)$ features in our feature pool.

5. Experiments

5.1. Experimental Setup

Datasets. Although there exist many face image benchmark datasets (e.g., LFW [17]), only a few of them are devoted to the study of age progression. We use the two most popular ones, the FG-Net dataset [6] and the MORPH [29] dataset, in our experiments. Both datasets have been widely used in age-related facial image analysis [4, 20, 22, 38]. Some example image pairs are shown in Fig. 2.

FG-Net [6] is the first popular face aging dataset and has been widely used for evaluating age-related facial image analysis tasks. The dataset contains 1002 images from 82 subjects, and the images are collected at ages in the range of 0 to 69. MORPH [29] is a large dataset containing two sections, MORPH Album I and MORPH Album II. Since Album I is small (only $1,690$ face images), in our experiments we use Album II, as suggested in [22], which has $55,132$ facial images of $13,617$ subjects.

Evaluation criterion. Following previous studies [20, 22, 24, 37, 38], we use *equal error rate* (EER), *i.e.*, the rate at which both accept and reject errors agree, for evaluating algorithm performance.

In addition, for better understanding the proposed algorithm, when comparing AGCD with different configurations (*e.g.*, $\lambda = 0$), we plot the performance curves for various numbers of iterations. Since different boosting algorithms learn at a different rate and begin over-fitting at different rounds, neither comparing results round by round or in the final round is not appropriate. In [39] it is suggested that the best test error over all rounds is more appropriate. Following the idea, we plot the *cumulative minimal EER* (CME) vs. iteration curves to study the effectiveness of the proposed algorithm (*e.g.*, Fig. 3(b)). A CME at a certain number iterations means the minimal EER value ever achieved before reaching the number of iterations.

Parameters. There are mainly two parameters in our algorithm, stepsize ϵ for hypothesis coefficient updating, and weight λ for the regularization term. We fix ϵ as 0.1 in all experiments by following the empirical and theoretical conclusion that small stepsizes often help in boosting [8, 42]. For λ, we fix it as 0.05 for when comparing with state-of-the-arts (Sec. 5.2). Analysis experiments are designed to evaluate different λ's (Sec. 5.3). For age verification, the age difference thresholds are automatically estimated as the median age differences of the training set. We use $1,500$ iterations for FG-net and $2,500$ iterations for MORPH, since they are large enough to reach the best performances for all algorithms according to the training process respectively (see Fig. 3(a) and Fig. 4(a)).

Table 1. Comparison with state-of-the-arts on FG-net. Results are quoted from the corresponding references except for "proposed".

Method	Year	EER (%)
Graph Matching [24]	2010	25.4
GOP [20]	2010	24.1
Landmark [37]	2010	23.6
Growth Model [38]	2012	22.3
AGCD (GOP)	proposed	21.7
AGCD	proposed	**19.4**

Table 2. Comparison with state-of-the-arts on MORPH. Results are quoted from [22] except for "proposed".

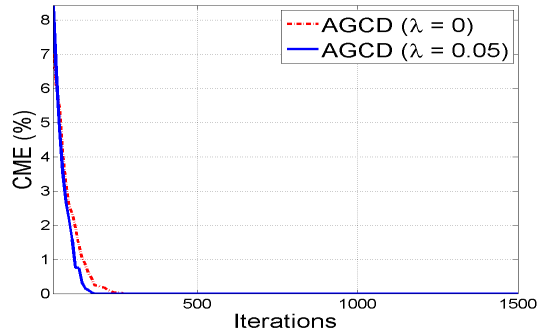
Method	Year	EER (%)
Bayesian Eigenface [28]	2005	9.7
GOP [20]	2010	10.5
Bagging LDA [18]	2011	10.2
NRML [22]	2014	8.6
MNRML [22]	2014	7.5
AGCD (GOP)	proposed	9.2
AGCD	proposed	**5.5**

5.2. Comparison with State-of-the-Arts

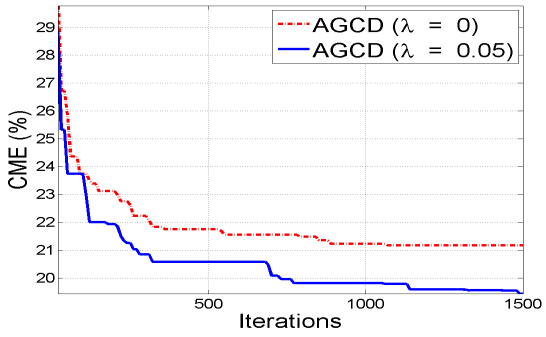
We first compare the proposed approach with the state-of-the-art cross-age face verification algorithms including *Bayesian Eigenface* [28], *Graph Matching* [24], *GOP* [20], *Landmark* [37], *Bagging LDA* [18], *Growth Model* [38], *NRML* [22] and *MNRML* [22]. These methods have previously tested on at least one of the two datasets in the same experimental settings. In addition these methods, we also run a version of AGCD using only the GOP features, so it can be compared clearly with GOP in [20].

The results are summarized in Tables 1 and 2, showing that the proposed algorithm clearly outperforms previous arts. Details of the experiments are given below.

FG-Net. We follow the experimental protocol used in [20, 24, 37, 38]. Specifically, the adult subset of FG-Net is used, which contains 272 images from 62 subjects. These images are used to generate 665 intra-personal pairs. Extra-personal pairs are randomly selected from images from different subjects. Three-fold cross validation is used, and there is no identity overlapping between different folds.

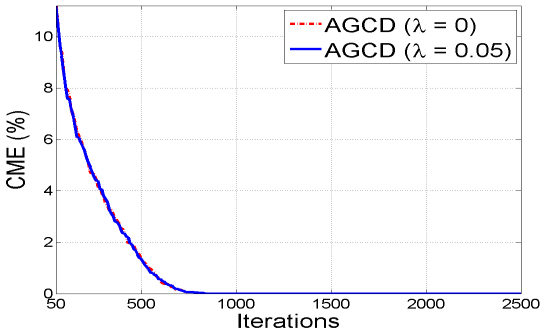


(a) Mean training CME

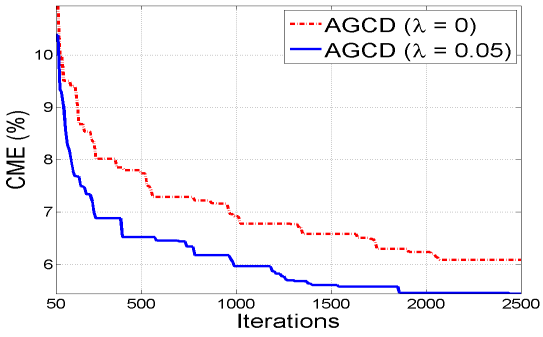


(b) Mean testing CME

Figure 3. Performances versus iterations on FG-Net.



(a) Mean training CME



(b) Mean testing CME

Figure 4. Performances versus iterations on MORPH.

Each fold contains about 220 intra-personal pairs and $2,000$ extra-personal pairs. Therefore, in each leave-one-fold-out round, there are approximately 440 intra-personal pairs (positive) and $4,000$ extra-personal (negative) pairs. For age verification, the proportion of positive and negative pairs are sampled to be similar for the face verification as discussed in Sec. 4.1. In particular, for the three training rounds, there are 439, 315, 355 positive and 1284, 1039, 1182 negative examples, respectively.

Table 1 summarizes the performances of the proposed algorithm along with other state-of-the-arts. It shows that our algorithm achieves an ERR of 19.4%, which significantly improves over the best published result (22.3%).

MORPH. We follow the experimental protocol in [22]: $13,000$ intra-personal pairs are generated by collecting image pairs for the same subjects with largest age gap. $15,000$ extra-personal pairs are randomly selected from images of different subjects. Three-fold cross validation was used, without subject overlapping between folds. Therefore, there are about $4,333$ intra-personal pairs and $5,000$ extra-personal pairs in each fold. The training sets for age verification are sampled in a similar way as for FG-Net, resulting 4821, 4795, 4700 positive samples, and 5179, 5205, 5300 negative samples for the three experimental rounds respectively.

The experimental results on MORPH are summarized in Table 2, which shows that our method again performs the best among all the algorithms. The best previously published result is 7.5%, while our algorithm achieves a better EER of 5.5%.

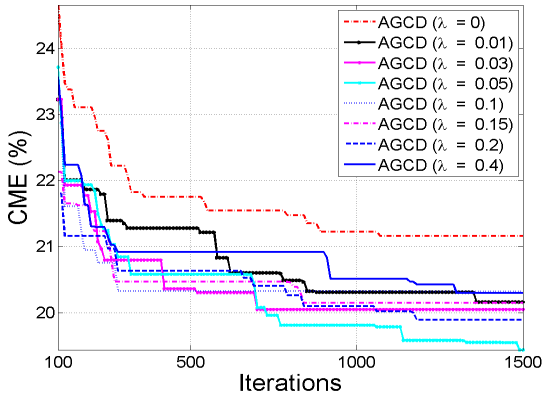
5.3. Parameter Analysis

We have also conducted experiments for further understanding the proposed algorithm from three aspects: behavior over iterations, effects of different λ's, and effects of feature sharing/competing.

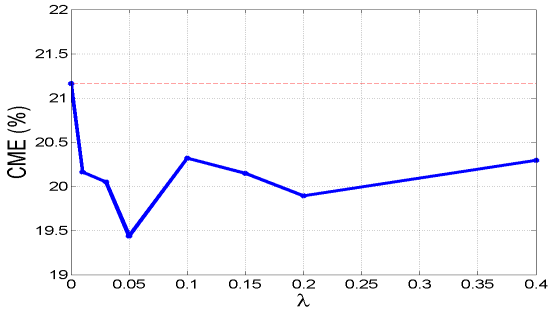
Behavior over number of iterations. The classification behavior over different numbers of iterations is important in analyzing boosting-like algorithms. We output CME-versus-iteration curves for two versions of the proposed algorithms, with $\lambda = 0.05$ (used when comparing with state-of-the-arts) and $\lambda = 0$ (degenerated to AdaBoost), respectively. The curves are plotted in Fig. 3 for FG-Net and Fig. 4 for MORPH.

The curves show that, while the both algorithms behave similarly during training, AGCD does improve the performance in testing. The observation implies that exploitation of task competition helps reducing overfitting.

Effect of λ. The weight λ controls the extent to which we regularize the proposed algorithm with the coordination constraints. For the analysis, we run the algorithm with different λ's taken from the set $\{0, 0.01, 0.03, 0.05, 0.1, 0.15, 0.2, 0.4\}$. The experiment is

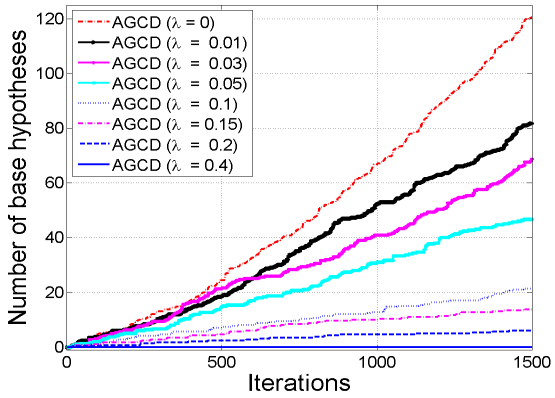


(a) CME versus iterations for different λ.

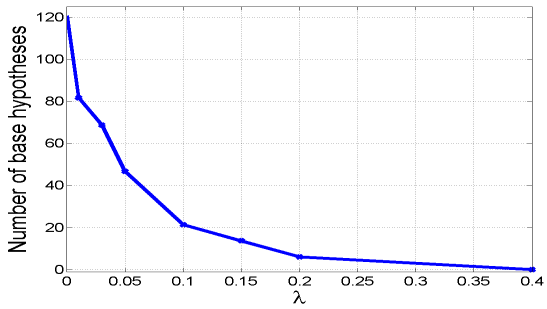


(b) CME of different λ for 1,500 iterations.

Figure 5. Performance (cumulative minimal EER (CME)) over parameter λ. This figure is best viewed in color.



(a) Number of base hypotheses with shared features versus iteration for different λ.



(b) Number of base hypotheses with shared features for different λ with 1,500 iterations.

Figure 6. Feature sharing in the proposed algorithm. This figure is best viewed in color.

conducted on FG-Net and the results are shown in Fig. 5. From the figures we can see that the best result is achieved when $\lambda = 0.05$. In addition, the results show that coordinating with the auxiliary task with a nonzero λ is consistently beneficial when compared with the baseline AdaBoost (*i.e.*, $\lambda = 0$).

Feature exclusion and sharing. The main motivation in this study is to boost visual recognition with task coordination, which, though encourages feature exclusion, allows as well feature sharing. Intuitively, a large λ encourages more feature exclusion, while a small λ is more comfortable with feature sharing. To investigate this phenomenon, we calculate the numbers of base hypotheses with common features by f and f_a in AGCD for different λ's and iterations. The results are plotted in Fig. 6, which confirms the above intuition. More importantly, the figure shows that the best performance is achieved when feature exclusion and feature sharing are well balanced ($\lambda = 0.05$).

Effects of different task weghts. In (3), the loss functions for target and auxiliary tasks are equally weighted. To exam the effects of different weights of the two tasks, we add a balancing weight parameter β to (3). The new formulation is shown below,

$$\mathcal{J}(\mathcal{D}, \mathcal{A}, \mathbf{p}, \mathbf{a}) = \mathcal{L}(\mathcal{D}, \mathbf{p}) + \beta\mathcal{L}(\mathcal{A}, \mathbf{a}) + \lambda \mathbf{p}^\top M_c^\top M_a \mathbf{a} \,. \quad (8)$$

Table 3. Effects of different weight β when $\lambda = 0.05$.

β	0.8	0.9	1.0	1.1	1.2
EER (%)	20.1	19.7	**19.4**	20.1	19.6

Table 3 shows the results of different β on FG-Net. From the table, we can see that empirically equally weighted task loss performs best.

6. Conclusion

In this paper, we have shown that exploiting the competing factors between cross-age face verification and age inference can be beneficial for improving cross-age face verification. Specifically, orthogonal regularization is incorporated into a greedy coordinate descent framework to derive a novel cross-age face verification algorithm by coordinating with age verification. The effectiveness of our algorithm is clearly demonstrated through experiments on the widely tested FG-Net and MORPH datasets, on both of them our algorithm outperforms all previously tested ones.

Acknowledgments We thank the anonymous reviewers for constructive comments. The work is supported in part by US NSF Grants IIS-1218156 and IIS-1350521.

References

[1] A. Asthana, S. Zafeiriou, S. Cheng and M. Pantic. Robust discriminative response map fitting with constrained local models. In *CVPR*, 2013. 5

[2] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *PAMI*, 28(12):2037–2041, 2006. 5

[3] M. Bereta, P. Karczmarek, W. Pedrycz, and M. Reformat. Local descriptors in application to the aging problem in face recognition. *PR*, 46(10):2634–2646, 2013. 2

[4] B. Chen, C. Chen, and W. Hsu. Cross-age reference coding for age-invariant face recognition and retrieval. In *ECCV*, 2014. 2, 6

[5] L. Du, and H. Ling. Exploiting competition relationship for robust visual recognition. In *AAAI*, 2014. 3

[6] FG-Net. Face and gesture recognition working group, 2000. [Online]. Available: http://www-prima.inrialpes.fr/FGnet/ 2, 6

[7] Y. Freund and R. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *J. of Comp. and Sys. Sci.*, 55(1):119–139, 1997. 3

[8] J. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232, 2001. 3, 5, 6

[9] Y. Fu, G. Guo, and T. Huang, Age synthesis and estimation via faces: a survey. *PAMI*, 32(11):1955–1976, 2010. 2

[10] D. Gong, Z. Li, D. Lin, J. Liu, and X .Tang. Hidden factor analysis for age invariant face recognition. In *ICCV*, 2013. 2

[11] P. Gong, J. Ye, and C. Zhang. Multi-stage multi-task feature learning. *JMLR*,14(10):2979-3010, 2013. 2

[12] G. Guo, G. Mu, and K. Ricanek. Cross-Age Face recognition on a very large database: the performance versus age intervals and improvement using soft biometric traits. In *ICPR*, 2010. 2

[13] G. Guo, G. Mu, Y. Fu, and T. Huang. Human age estimation using bio-inspired features. In *CVPR*, 2009. 5, 6

[14] G. Guo and G. Mu. Joint estimation of age, gender and ethnicity: CCA vs. PLS. In *FG*, 2013. 3

[15] H. Han, C. Otto, X. Liu, and A. Jain. Demographic estimation from face images: human vs. machine performance. *PAMI*, Sep. 2014 (In press). 5

[16] S. Hwang, K. Grauman and F. Sha. Learning a tree of metrics with disjoint visual features. In *NIPS*, 2011. 3

[17] G. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: a database for studying face recognition in unconstrained environments. University of Massachusetts, Amherst, Technical Report 07-49, October, 2007. 6

[18] Z. Li, U. Park, and A. Jain. A discriminative model for age invariant face recognition. *IEEE T-IFS*, 6(3):1028–1037, 2011. 2, 5, 6

[19] H. Ling, S. Soatto, N. Ramanathan, and D. Jacobs. A study of face recognition as people age. In *ICCV*, 2007. 1

[20] H. Ling, S. Soatto, N. Ramanathan, and D. Jacobs. Face verification across-age progression using discriminative methods. *IEEE T-IFS*, 5(1):82–91, 2010. 2, 5, 6

[21] P. Liu, T. Zhou, W. Tsang, Z. Meng, S. Han, and Y. Tong. Feature disentangling machine - a novel approach of feature selection and disentangling in facial expression analysis. In *ECCV*, 2014. 3

[22] J. Lu, X. Zhou, Y. Tan, Y. Shang, and J. Zhou. Neighborhood repulsed metric learning for kinship verification. *PAMI*, 36(2):331–345, 2014. 6, 7

[23] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91-110, 2004. 5

[24] G. Mahalingam and C. Kambhamettu. Age invariant face recognition using graph matching. In *BTAS*, 2010. 6

[25] S. Pan and Q. Yang. A survey on transfer learning. *IEEE T-KDE*, 22(10):1345–1359, 2010. 2

[26] U. Park, Y. Tong, and A. Jain. Age-invariant face recognition. *PAMI*, 32(5):947–954, 2010. 2

[27] N. Ramanathan and R. Chellappa. Modeling age progression in young faces. In *CVPR*, 2006. 2

[28] N. Ramanathan and R. Chellappa. Face verification across-age progression. In *CVPR*, 2005. 2, 6

[29] K. Ricanek and T. Tesafaye. MORPH: a longitudinal image database of normal adult age-progression. In *FG*, 2006. 2, 6

[30] M. Riesenhuber and T. Poggio. Hierarchical models of object recognition in cortex. *Nature Neuroscience*,2(11):1019-1025, 1999. 5, 6

[31] B. Romera-Paredes, A. Argyriou, N. Berthouze, and M. Pontil. Exploiting unrelated tasks in multi-task learning. In *AISTATS*, 2012. 3

[32] Y. Sun, X. Wang, and X. Tang. Hybrid Deep Learning for Face Verification. In *ICCV*, 2013. 2

[33] D. Sungatullina, J. Lu, G. Wang, and P. Moulin. Multiview discriminative learning for age-invariant face recognition. In *FG*, 2013. 2, 5

[34] J. Suo, S. Zhu, S. Shan, and X. Chen. A compositional and dynamic model for face aging. *PAMI*, 32(3):385–401, 2010. 2

[35] A. Torralba, K. Murphy, and W. Freeman. Sharing features: efficient boosting procedures for multiclass object detection. In *CVPR*, 2004. 3

[36] J. Wang, Y. Shang, G. Su, and X. Lin. Age simulation for face recognition. In *ICPR*, 2006. 2

[37] T. Wu, P. Turaga, and R. Chellappa. Age estimation and face verification across aging using landmarks. *IEEE T-IFS*, 7(6):1780-1788, 2012. 6

[38] T. Wu and R. Chellappa. Age invariant face verification with relative craniofacial growth model. In *ECCV*, 2012. 6

[39] Y. Xi, Z. Xiang, P. Ramadge, and R. Schapire. Speed and sparsity of regularized boosting. In *AISTATS*, 2009. 6

[40] Z. Xiang, Y. Xi, U. Hasson, and P. Ramadge. Boosting with spatial regularization. In *NIPS*, 2009. 4

[41] X. Yuan and S. Yan. Visual classification with multitask joint sparse representation. In *CVPR*, 2010. 3

[42] T. Zhang and B. Yu. Boosting with early stopping: Convergence and consistency. *Annals of Statistics*, 33(4):1538-1579, 2005. 5, 6

[43] W. Zhao, R. Chellappa, P. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Comput. Surv.* 35(4):399-458,2003. 2

[44] D. Zhou, L. Xiao, and M. Wu. Hierarchical classification via orthogonal transfer. In *ICML*, 2011. 3

[45] Y. Zhou, R. Jin, and S. Hoi. Exclusive lasso for multitask feature selection. In *AISTATS*, 2010. 3