Demographic Estimation from Face Images: Human vs. Machine Performance

Hu Han, *Member, IEEE,* Charles Otto, *Student Member, IEEE,* Xiaoming Liu, *Member, IEEE* and Anil K. Jain, *Fellow, IEEE*

Abstract—Demographic estimation entails automatic estimation of age, gender and race of a person from his face image, which has many potential applications ranging from forensics to social media. Automatic demographic estimation, particularly age estimation, remains a challenging problem because persons belonging to the same demographic group can be vastly different in their facial appearances due to intrinsic and extrinsic factors. In this paper, we present a generic framework for automatic demographic (age, gender and race) estimation. Given a face image, we first extract demographic informative features via a boosting algorithm, and then employ a hierarchical approach consisting of between-group classification, and within-group regression. Quality assessment is also developed to identify low-quality face images that are difficult to obtain reliable demographic estimates. Experimental results on a diverse set of face image databases, FG-NET ($1K$ images), FERET ($3K$ images), MORPH II ($75K$ images), PCSO ($100K$ images), and a subset of LFW ($4K$ images), show that the proposed approach has superior performance compared to the state of the art. Finally, we use crowdsourcing to study the human perception ability of estimating demographics from face images. A side-by-side comparison of the demographic estimates from crowdsourced data and the proposed algorithm provides a number of insights into this challenging problem.

Index Terms—Demographic estimation, demographic informative feature, quality assessment, hierarchical approach, crowdsourcing, human vs. machine

1 Introduction

Humans can glean a wide variety of information from a face image, including identity, age, gender, and race (see Fig. 1). The identification-specific characteristics of face images have been well explored in face recognition research and various applications [1], e.g., access control, video surveillance, and criminal investigation. In contrast, there is relatively less research [2] on how to accurately estimate the *demographic information* from a face image. Specifically, we consider age, gender, and race in this paper.

There has been a growing interest in automatic extraction of demographic information from face images or videos, due to many emerging applications [2], [5], [6]. These include (i) *access control*, e.g., an automatic age estimation system can prevent minors from purchasing alcohol or cigarette from vending machines; (ii) *human-computer interaction*, e.g., a smart shopping cart can dynamically change advertisement on a bill board based on the demographics of the customers passing by; and (iii) *law enforcement*, e.g., an automatic demographic estimation system can help to identify the suspect more efficiently and accurately by filtering the mugshot database with the estimated age, gender, and race from the face image of a suspect.

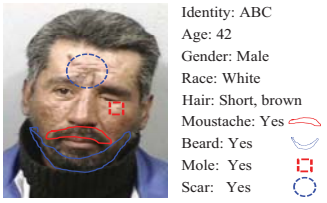


Fig. 1. A wide variety of information can be gleaned from a face image, such as identity, age, gender, race, and scars, marks and tattoos (SMT).

Despite recent progress [7], [8], [33], automatic demographic estimation remains a difficult problem. The challenges in automatic demographic estimation come from the large intra-class facial appearance variations due to both intrinsic and extrinsic factors[1]. Figure 3 shows that extrinsic factors, such as environment, lifestyle, and health, could lead to dramatically different facial appearances of identical twins.

1.1 Proposed Approach

Among the sizable literature on demographic estimation summarized in Tables 1–3, most prior work is limited to estimating a *single* demographic attribute. Research on age, gender, and race estimation via a generic framework is still quite limited. For example, only three publications [7], [8], [33] provide a joint

• *Hu Han, Charles Otto, Xiaoming Liu and Anil K. Jain are with the Department of Computer Science and Engineering, Michigan State University, East Lansing, MI 48824, USA.*

E-mail: {hhan, ottochar, liuxm, jain}@msu.edu

1. http://www.skincarephysicians.com/agingskinnet/basicfacts.html

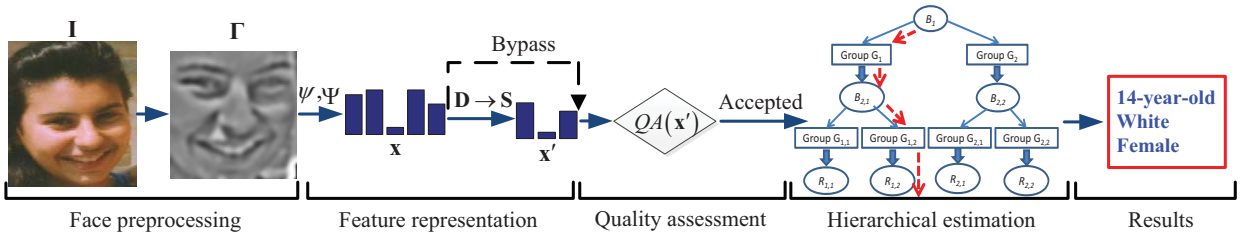
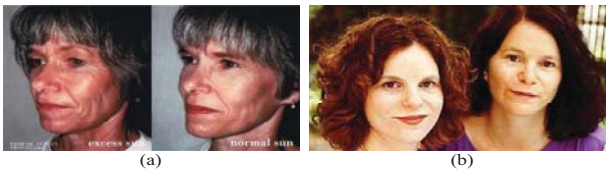


Fig. 2. Overview of the proposed approach for automatic demographic estimation. $\mathbf{I}$ and $\mathbf{\Gamma}$ are the input and preprocessed face images, respectively; ψ and Ψ denote the BIF extraction process, and $\mathbf{D} \rightarrow \mathbf{S}$ denotes the demographic informative feature selection; $\mathbf{x}$ and $\mathbf{x}'$ denote the BIF feature vector before and after feature selection, respectively; and $QA(\cdot)$ is the quality assessment process.



(a) (b)

Fig. 3. Different facial appearances of identical twins possibly due to extrinsic factors such as (a) environmental conditions (e.g., sunshine)[2], and (b) lifestyle[3].

estimate of age, gender, and race using a generic framework. Even these methods have some serious limitations. For example, age estimation in [7] is limited to age group classification, namely child, youth and old. Approaches in [8], [33] have fairly high computational costs. Additionally, studies on human perception of demographics from face images are still sparse. For example, the largest datasets used to obtain human estimates of age and gender only contain 500 and 300 images, respectively (see Tables 1 and 2). For race classification, to our knowledge, no performance on human estimates has been reported (see Table 3).

In this paper, we present a generic framework for automatic demographic estimation from a single face image (see Fig. 2). We extract the previously proposed biologically inspired features (BIF) [24] from a face image, and select demographic informative features using a boosting algorithm. We then propose a hierarchical estimator consisting of between-group classification and within-group regression to predict the age, gender, and race. We also design a quality assessment method to detect low-quality face images, which arise from large pose, illumination, and expression variations. Further, we report human perceptual ability in demographic estimation using crowdsourcing on a diverse set of face image databases ($1,002$ images from FG-NET [20], $2,000$ images from MORPH II [63], and $4,200$ images from the Pinellas County Sheriff's Office (PCSO)) described in Appendix A. This allows a comparison of the abilities of machine and human to estimate demographics.

The main contributions of this paper include: (i) a generic framework for demographic estimation (age, gender, and race) from a face image; (ii) a hierarchical approach for coarse-to-fine age estimation; (iii) a face image quality assessment method for the purpose of demographic estimation; (iv) studies on the ability of human in demographic estimation using crowdsourcing; and (v) studies on the generalization ability of the proposed approach with cross-database testing, and data drawn from the general population of faces.

This paper is built upon our preliminary work [56]. The main differences are summarized as follows. (i) While [56] focused only on age estimation, this work addresses age, gender, and race estimation. (ii) Demographic informative features are designed using BIF with a boosting algorithm. (iii) A quality assessment metric is introduced to enable rejection of low-quality face images for the purpose of demographic estimation. (iv) Human estimates on gender and race are studied. (v) Comprehensive evaluations are performed using the FG-NET, FERET, MORPH II, PCSO, and LFW databases.

1.2 Prior Work

A number of studies in the biological, psychological, and cognitive sciences have reported on how the human brain perceives, represents, and remembers faces. In particular, various aspects of human demographic estimation have been studied in the field of psychology [9]–[11]. These studies provide some context for the performance of automatic demographic estimation reported in the literature. However, to our knowledge, no *large scale* studies on the accuracy of human demographic estimation have been reported on public-domain databases (e.g., MORPH II [63] and FG-NET [20]). On the related note, there are some studies on human vs. machine performance on face recognition. Due to limited space, we refer interested readers to reviews in [3], [4].

Attempts to design computational models based on psychological studies for automatic demographic estimation started in the 1990s [12], [13]. Since then,

2. http://www.antell-md.com/newyorkplasticsurgeon/plasticsurgerytwins.htm

3. http://www.cbc.ca/news/background/health/identical-twins.html

TABLE 1
A summary of machine and human performance on age estimation reported in literature.

Publication	Face representation	Face database (#subjects, #images)	Human perception of age	Performance measure and accuracy (MAE†, 5-year CS‡)
Lanitis *et al.* (2002) [17]	2D shape, Raw intensity	FG-NET (NA, 565)	Studied on 32 face images	MAE: 4.3
Geng *et al.* (2007) [19]	Active appearance model (AAM)	FG-NET (82, 1002) MORPH (NA, 433)	Studied on 51 face images from FG-NET	FG-NET / MORPH MAE: 6.8 / 8.8 CS: ∼65% / ∼46%
Yang and Ai (2007) [7]	LBP, Haar-like features	FERET (1196, 3540) CMU-PIE (68, 696) Snapshot (NA, 9000)	Not studied	FERET / CMU-PIE / Snapshot Age group: 92.1% / 87.5% / 93.2%
Fu and Huang (2008) [21]	Manifold of raw intensity	YGA (1600, 8000)	Not studied	MAE: 5∼6 CS: F: ∼55%, M: ∼50%
Suo *et al.* (2008) [22]	Holistic and local topology, 2D shape, color, and gradient	FG-NET (82, 1002) Private (NA, 8000)	Studied on 500 images from a private database	FG-NET / Private MAE: 6.0 / 4.7 CS: ∼55% / ∼66%
Guo *et al.* (2009) [23]	Biologically inspired features (BIF)	FG-NET (82, 1002) YGA (1600, 8000)	Not studied	FG-NET / YGA MAE: 4.8 / F: 3.9, M: 3.5 CS: 47% / F: 75%, M: 80%
Guo and Wang (2011) [25]	BIF	MORPH II (NA, 55132)	Not studied	MAE: 4.2
Choi *et al.* (2011) [26]	AAM, Gabor, LBP	FG-NET (82, 1002) PAL (NA, 430) BERC (NA, 390)	Not studied	FG-NET / PAL / BERC MAE: 4.7 / 4.3 / 4.7 CS: ∼73% / ∼70% / ∼65%
Luu *et al.* (2011) [27]	Holistic contourlet appearance model	FG-NET (82, 1002) PAL (NA, 443)	Not studied	FG-NET / PAL MAE: 4.1 / 6.0 CS: ∼74% / NA
Chang *et al.* (2011) [28]	AAM	FG-NET (82, 1002) MORPH II (NA, 5492)	Not studied	FG-NET / MORPH II MAE: 4.5 / 6.1 CS: 74.4% / 56.3%
Wu *et al.* (2012) [29]	Grassmann manifold of 2D shape	FG-NET (82, 1002) Passport (109, 233)	Not studied	FG-NET / Passport MAE: 5.9 / 8.8 CS: 62% / 40%
Thukral *et al.* (2012) [30]	Grassmann manifold of 2D shape	FG-NET (82, 1002)	Not studied	MAE: 6.2
Chao *et al.* (2013) [32]	AAM with distance metric adjustment	FG-NET (82, 1002)	Not studied	MAE: 4.4
Lu and Tan (2013) [31]	Manifold of raw intensity	MORPH II (NA, 20000)	Not studied	White / Black MAE: 5.2 / 4.2 CS: 67% / 59%
Hadid and Pietikäinen (2013) [8]	Raw intensity, volume LBP	Internet videos (NA, 2000)	Not studied	Age group classification: 83.1%
Guo and Mu (2013) [33]	BIF	MORPH II (NA, 55132)	Not studied	MAE: 4.0
Geng *et al.* (2013) [34]	AAM, BIF	FG-NET (82,1002) MORPH II (13000, 55132)	Studied on 51 and 60 images, respectively, from FG-NET and MORPH II	FG-NET / MORPH II MAE: 4.8 / 4.8
Proposed method	Demographic informative features	FG-NET (82, 1002) MORPH II (20569, 78207) PCSO (81533, 100012) LFW (4211, 4211)	Studied on 1002, 2000, 2200 and 4211 images, respectively, from FG-NET, MORPH II, PCSO, and LFW	FG-NET / MORPH II / PCSO / LFW MAE: 3.8 / 3.6 / 4.1 / 7.8 CS: 78.0% / 77.4% / 72.6% / 42.5%

†MAE (Mean Absolute Error) [17] is the average of the absolute difference between the estimated age and ground-truth age (true age). ‡CS (Cumulative Score) [19] reflects the percentage of correct age estimates w.r.t. different absolute errors.

a significant progress has been made on automatic demographic estimation, and a number of approaches have been reported in the literature [8], [17], [19], [21], [23], [27], [29], [33], [39], [44], [51], [54].

A few survey papers on demographic estimation methods are available. Most of the age estimation methods published prior to 2011 are reviewed in [16]. A comparison of representative gender classification methods is provided by Mäkinen and Raisamo [35]. Here, we provide a brief review of existing demographic estimation methods by grouping them into three main categories: anthropometry-based approach, image-based approach, and appearance-based approach.

The anthropometry-based approach adopted in [13], [14], [22], [36], utilizes the distance ratios between individual facial landmarks to describe the topological and configural differences among face shapes across different demographic groups. In age estimation, it is natural to consider the anthropometric features due to the craniofacial growth; however, they are mainly useful to distinguish children from adults, since the facial shape becomes quite stable for adults [21]. Additionally, anthropometry-based approach requires accurate facial landmark localization and, in some cases, even manual annotation, which limits their usability in automatic demographic estimation systems.

Image-based approach, such as in [7], [18], [23], [25], [33], [39], [54], differentiates faces of different demographic groups by relying on texture, e.g., skin, wrinkle and facial marks. Texture features such as Gabor, LBP, PCA, Haar, and BIF have been widely used to represent both the holistic and local face regions. Image-based approach has been found to be effective in all the three demographic estimation tasks considered here, but it is typically less efficient due to high feature dimensionality.

The appearance-based approach, such as in [17], [19], [21], [26], [27], [34], [43], [48], utilizes facial appearance (both texture and shape) to differentiate faces among individual demographic groups. Active Appearance Model (AAM) and its variations [27], [42] are widely used to model the facial texture and shape. Similar to the anthropometry-based approach, the appearance-based approach demands highly accurate facial landmark localization.

In Tables 1–3, we summarize machine and human performance on age, gender and race estimation tasks, respectively, covering automatic estimation approaches, human estimation experiments, databases, and performance. Public-domain face databases that are widely used in such studies are described in Appendix A.

The remainder of this paper is structured as follows. In Section 2 we detail the proposed automatic demographic estimation approach. In Section 3 we describe

TABLE 2
A summary of machine and human performance on gender classification reported in literature.

Publication	Face representation	Face database (#male, #female) images	Human perception of gender	Gender classification accuracy
Moghaddam and Yang (2000) [37]	Raw intensity	FERET (1044, 711)	Studied on 254 face images	96.6%
Gutta *et al.* (2000) [39]	Raw intensity	FERET (1906, 1100)	Not studied	96%
Wu *et al.* (2003) [41]	Haar-like features	FERET (5500, 5500) Private (1300, 1300)	Not studied	88%
Saatci and Town (2006) [43]	AAM	Combined (NA, NA)	Not studied	94.8%
Baluja and Rowley (2007) [44]	Pixel intensity difference	FERET (1495, 914)	Not studied	94.4%
Yang and Ai (2007) [7]	LBP, Haar-like features	FERET (NA, NA) CMU-PIE (NA, NA) Snapshot (4696, 3737)	Not studied	FERET / CMU-PIE / Snapshot 93.3% / 91.1% / 96.3%
Gao and Ai (2009) [45]	Raw intensity	Snapshot (900, 900) Consumer (650, 650) Multi-race (1200, 1200)	Not studied	Snapshot / Consumer 93.7% / 92.8% Multi-race (Mong./Cau./Afr.) 91.5% / 92.8% / 88%
Tariq *et al.* (2009) [55]	2D shape	Silhouetted profile (230, 211)	Not studied	71.2%
Wu *et al.* (2010) [46]	Facial surface normals	UND (100, 100) FERET (100, 100)	Studied on 80 face images	UND / FERET 91.7% / 83.4%
Mozaffari *et al.* (2010) [47]	LBP, DCT, and geometric distances	AR (56, 70) Iranian (56, 70)	Not studied	AR / Iranian 96.0% / 97.1%
Wang *et al.* (2010) [48]	AAM and LPP	FG-NET (555, 447)	Studied on 8 face images	84.3%
Dong and Woodard (2011) [49]	Eyebrow shape	MBGC (59, 99) FRGC (50, 50)	Not studied	MBGC / FRGC 96% / 97%
Zhang and Wang (2011) [50]	Hierarchical and discriminative Bag-of-Words	UND Collection F (562, 380)	Not studied	97.7%
Bekios-Calfa *et al.* (2011) [51]	PCA and ICA	UCN (5628, 5041) FERET (591, 402) PAL (219, 357)	Not studied	UCN / FERET / PAL 95.4% / 94.0% / 89.8%
Ballihi *et al.* (2012) [52]	3D geometrical features	FRGC v2 (264, 202)	Not studied	86.1%
Tapia and Perez (2013) [54]	LBP	FERET (212, 199) UND (301, 186) LFW (4500, 2943)	Not studied	FERET / UND / LFW 99.1% / 94.0% / 98.0%
Chen and Ross (2013) [53]	Local gradient Gabor pattern	AR (50, 50)	Not studied	94%
Hadid and Pietikäinen (2013) [8]	Raw intensity, volume LBP	Combined (NA, 1000)	Not studied	96.8%
Guo and Mu (2013) [33]	BIF	MORPH II (46645, 8487)	Not studied	96.0%
Proposed method	Demographic informative features	MORPH II (65601, 12606) PCSO (75006, 25006) FERET (1722, 1007) LFW (4211, 4211)	Studied on 2000, 2200, and 4211 images, respectively, from MORPH II, PCSO and LFW databases	MORPH II / PCSO / FERET / LFW 97.6% / 97.1% / 96.8% / 94%

TABLE 3
A summary of machine and human performance on race classification reported in literature.

Publication	Face representation	Face race database (#races, #images)	Human perception of race	Race classification accuracy
Gutta *et al.* (2000) [39]	Raw intensity	FERET (4, 3006)	Not studied	92%
Yang and Ai (2007) [7]	LBP, Haar-like features	CMU-PIE (2, 696) Merged (2, 12696)	Not studied	CMU-PIE / Merged 93.2% / 97.0%
Tariq *et al.* (2009) [55]	Shape	Profile (4, 441)	Not studied	71.7%
Hadid and Pietikäinen (2013) [8]	Raw intensity, volume LBP	Videos (2, NA)	Not studied	100%
Chen and Ross (2013) [53]	Local gradient Gabor pattern	MORPH and CAS-PEAL (3, 750)	Not studied	98.7%
Guo and Mu (2013) [33]	BIF	MORPH II (2, 53160)	Not studied	98.9%
Proposed method	Demographic informative features	MORPH II (2, 78207) PCSO (2, 100012) LFW (2, 4211)	Studied on 2000 images each from MORPH II and PCSO databases, and 4211 images from LFW database	MORPH II / PCSO / LFW 99.1% / 98.7% / 90%

While [39], [55], and [53] studied race classification using four (Caucasian, South Asian, East Asian, and African) and three (Asian, Caucasian, and African) race groups, respectively, no human performance is reported on these race groups. In addition, black and white groups constitute the majority of subjects in the public-domain MORPH II database (see Appendix A). Therefore, in this work, we focus on the classification between black and white.

the experiments on demographic estimation by humans using crowdsourcing. Experimental results are presented in Section 4, and finally we conclude this work in Section 5.

2 Automatic Demographic Estimation

2.1 Face Preprocessing

There are many different types of appearance variations in facial images. For example, as shown Fig. 4 (a), face images can be either in gray-scale or color, and some of the color images have a color cast. Therefore, color in face images is not always available and reliable. There are also large pose variations, which lead to misalignment between face images.

To compensate for these variations, we design a face preprocessing procedure. We first convert a color face image into a gray-scale image to mitigate the influence of inconsistent colors. We utilize a widely used model in PAL and NTSC: $\mathbf{I} = 0.2989 * \mathbf{R} + 0.5870 * \mathbf{G} + 0.1140 * \mathbf{B}$, where $\mathbf{R}$, $\mathbf{G}$, and $\mathbf{B}$ are the red, green, and blue channels of a color image, respectively; $\mathbf{I}$ is the output gray-scale face image. To reduce the effect of scale, rotation, and translation variations, a face image is then rectified based on the two eyes and cropped to 60×60 pixels with a 32-pixel interpupillary distance (IPD). We detect the face and the eyes using Cognitec's FaceVACS SDK [57]. Finally, to suppress both low-frequency illumination variation and high-frequency noise (e.g., photon and sensor noise), we apply Difference of Gaussians (DoG) filtering,

$$\mathbf{\Gamma}(x, y) = \mathbf{I}(x, y) * (\mathbf{G}(x, y, \sigma_1) - \mathbf{G}(x, y, \sigma_2)), \quad (1)$$

where $\mathbf{G}(\cdot)$ is a Gaussian smoothing function. In our experiments, we use $\sigma_1 = 0.2$, and $\sigma_2 = 1.0$ for DoG filtering.

2.2 Demographic Informative Representation

Since we aim to perform age, gender, and race estimations under a generic framework, it is desirable to have a face representation appropriate for all these three tasks. Following the success of biologically inspired features (BIF) [58] in object detection and

Fig. 4. Examples of face preprocessing: (a) input images with pose variations from FG-NET and MORPH II; (b) intermediate gray-scale images after geometric normalization; and (c) output of preprocessing.

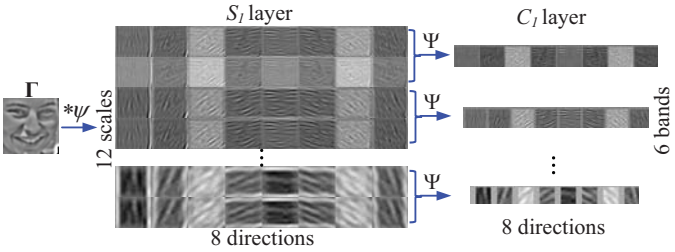


Fig. 5. Major steps in calculating biologically inspired features (BIF).

classification [59], face recognition [60], and automatic age estimation [23], [25], [33], we choose to use BIF.

In its simplest form, the extraction of BIF consists of two layers of computational units, where simple S_1 units in the first layer are followed by complex C_1 units in the second layer. The S_1 units correspond to the classical simple cells in the primary visual cortex [64]. They are typically implemented with the convolution of a preprocessed image $\mathbf{\Gamma}$ with a family of Gabor filters [61],

$$\psi_{u,v}(\mathbf{z}) = \frac{\|\mathbf{k}_{u,v}\|^2}{\sigma^2} e^{-\frac{\|\mathbf{k}_{u,v}\|^2 \|\mathbf{z}\|^2}{2\sigma^2}} \left[e^{i\mathbf{k}_{u,v}\mathbf{z}} - e^{-\frac{\sigma^2}{2}} \right], \quad (2)$$

where $\mathbf{z} = (x, y)$, σ is the relative width of the Gaussian envelope function w.r.t. the wavelength, and u and v are the orientation and scale parameters of Gabor kernels, respectively. The wave vector $\mathbf{k}_{u,v}$ is defined as,

$$\mathbf{k}_{u,v} = k_v e^{i\phi_u}, \quad (3)$$

with $k_v = \frac{k_{max}}{f^v}$ defining the frequency, and $\phi_u = \frac{\pi u}{8}$ defining the orientation. k_{max} and f are constants specifying the maximum frequency and scaling factor between two neighboring kernels, respectively. The C_1 units correspond to cortical complex cells which are robust to shift and scale variations. They can be calculated by pooling over the preceding S_1 units with the same orientation but at two successive scales.

To compute S_1 layer features, we build a family of Gabor filters similar to those in [61], but we use 8 orientations ($u \in [0, 7]$) and 12 scales ($v \in [1, 12]$) as suggested in [23]. We apply "MAX" pooling operator [24] and "STD" normalization operator [23] to extract C_1 features from the S_1 layer. The 8 orientation and 6 scale features in the C_1 layer are finally concatenated into a single feature vector $\mathbf{x}$. As shown in Fig. 5, the BIF extraction is denoted as,

$$\mathbf{x} = \Psi_{\text{MAX,STD}}(\text{Re}(\mathbf{\Gamma} * \psi)), \quad (4)$$

where Ψ denotes two consecutive operations of "MAX" and "STD" on a filtered image $\mathbf{\Gamma}$.

The S_1 layer provides a multi-scale representation for face images, and the C_1 layer provides robustness against translation, rotation, and scaling that could not be handled in the face preprocessing stage. However, a multi-scale representation also implies a high feature dimensionality. In our approach, the dimensionality d of the feature vector $\mathbf{x}$ is $4,280$, which makes it difficult to perform demographic estimation *efficiently*. To address this issue, based on a training set $\mathbf{D}$, we perform feature selection (FS) to reduce the feature dimensionality while retaining the discriminability of BIF features.

Formally, given a training set with m samples,

$$\mathbf{D} = \{(\mathbf{x}_i, y_i) : \mathbf{x}_i \in \mathbb{R}^d, y_i \in \mathbb{N}, i \in [1, m]\}, \quad (5)$$

where $\mathbf{x}_i$ is a d-dimensional BIF vector from the i^{th} training image, and y_i is its associated label, our objective is to select a low-dimensional (d') subset of BIF features,

$$\mathbf{S} = \{\mathbf{x}'_i : \mathbf{x}'_i \in \mathbb{R}^{d'}, \mathbf{x}'_i \subset \mathbf{x}_i, i \in [1, m]\}, \quad (6)$$

$d' \ll d$, that can retain the discriminative information for demographic estimation. In our experiments, $d' = 800$ is used for all the three demographic estimation tasks. The label y_i depends on individual demographic estimation tasks: for age estimation, $y_i \in [0, 70]$; for gender and race classification, y_i represents binary classes, e.g., $y_i \in \{0, 1\}$ indicating {male, female} or {white, black}.

We use multi-class AdaBoost [40], [65] to select the demographic informative features. A multi-level decision tree (DT) is used as the weak classifier $h_j(\mathbf{x}_i)$. In our experiments, we use a 7-level decision tree for age estimation, and 1-level decision tree for gender and race classification. The feature selection procedure is outlined in Algorithm 1.

Figure 6 shows the top 50 most informative features for the three different demographic estimation problems; the extracted features look fairly symmetric. The most informative features for age estimation are located in the regions where wrinkles typically appear [66], such as the eye and mouth corners, nasolabial folds, and cheeks. For gender classification, besides the features located around the eyes and lip, the jaw is also found to be salient, which is consistent with human perception experiments [67]. For race (white vs. black) classification, the most informative features are around the eyes, nose, and lip, which is also consistent with human perception studies [68].

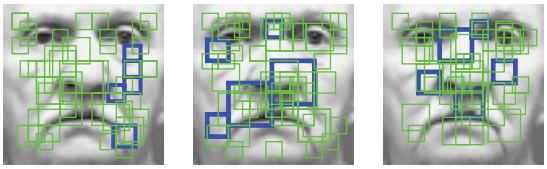
Input: Training set $\mathbf{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{m}$; desired feature dimensionality d'.
Output: Selected feature set $\mathbf{S}$.
Initialization: $w_i^{(1)} \leftarrow \frac{1}{m}$; $\mathfrak{J} \leftarrow \emptyset$
for $t \leftarrow 1$ **to** d' **do**
  Normalize the weights: $w_i^{(t)} \leftarrow \frac{w_i^{(t)}}{\sum_{j=1}^{m} w_i^{(t)}}$
  for $j \leftarrow 1$ **to** d *and* $j \notin \mathfrak{J}$ **do**
    Train a multi-class weak classifier, $h_j(\mathbf{x}_i)$
    Compute the weighted error:
    $\epsilon_j^{(t)} \leftarrow \sum_{i=1}^{m} w_i^{(t)} e\left(h_j(\mathbf{x}_i), y_i\right)$, where $e(\cdot) = 0$ if $\mathbf{x}_i$ is classified correctly and 1 otherwise
  Find the best feature index: $\hat{j} \leftarrow \underset{j}{\operatorname{argmin}}\, \epsilon_j^{(t)}$
  Update the selected indexes: $\mathfrak{J} \leftarrow \mathfrak{J} \cup \{\hat{j}\}$
  Update the weights:
  $w_i^{(t+1)} \leftarrow w_i^{(t)} \beta_t^{1-e(h_{\hat{j}}(\mathbf{x}_i), y_i)}$, where $\beta_t = \frac{\epsilon_{\hat{j}}^{(t)}}{1-\epsilon_{\hat{j}}^{(t)}}$
Result: $\mathbf{S} = \{\mathbf{x}_i(\mathfrak{J})\}_{i=1}^{m}$.

Algorithm 1: Demographic informative feature selection.

2.3 Hierarchical Demographic Estimation

While age estimation is naturally formulated as a regression problem, gender and race classifications are binary classification tasks. To handle regression and classification problems using a generic framework, we present a hierarchical approach consisting of a classification stage followed by a regression stage. In the classification stage (Fig. 7 (a)), three binary SVM classifiers (B_1, $B_{2,1}$ and $B_{2,2}$) are used to build a two-level binary decision tree (BDT) [62], and a test face image is classified into one of the four groups. In the regression stage (Fig. 7 (b)), a separate SVM regressor ($R_{1,1}$, $R_{1,2}$, $R_{2,1}$, and $R_{2,2}$) is trained within each group to make an accurate age prediction. With this hierarchical estimation approach, we can perform different demographic estimation tasks flexibly. For example, while age estimation goes through both the classification and regression stages, gender and race classifications only require the B_1 classifier in the classification stage. Thus, we have three different B_1 classifiers that are trained using age, gender, and race data, respectively. The output of B_1 is always binary (0 or 1). Additionally, hierarchical methods have also been found to be more effective than direct regressions for age estimation [26], [30].

Hierarchical methods have been studied for age estimation problems in [26], [30], but our approach differs from these methods in two aspects. (i) While [26], [30] directly partition the entire age range (e.g., 0–70) into multiple groups, we use a BDT to perform coarse-to-fine classification. (ii) After the group classification stage, we train individual regression models with overlapping age ranges (i.e., age overlap Δ in



(a) (b) (c)

Fig. 6. The top 5 (blue) and top 6-50 (green) most informative BIF features selected for estimating (a) age, (b) gender, and (c) race. The rectangle size indicates the scale of the corresponding BIF feature.

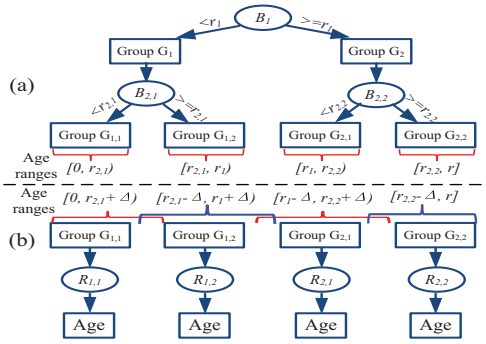


Fig. 7. Learning a hierarchical age estimator: (a) binary decision tree for classifying non-overlapping groups (e.g., male vs. female; white vs. black; age $< r_1$ vs. age $\geq r_1$), and (b) within-group age regressors learned from overlapping age groups.

Fig. 7 (b)). This mitigates the error due to incorrect age group classification of face images that are close to the group boundaries. Since we only use overlapping group ranges in the regression stage, this does not introduce label ambiguity [26] during the classification stage. Age estimation experiments on FG-NET show that the hierarchical estimation approach *without* using BDT or overlapping ranges leads to higher MAEs (5.0 and 5.2 years, respectively) than the proposed approach (4.8 years). We use the RBF kernel for all the SVM classifiers and regressors. For each dataset, we use $\Delta = 5$, and select the parameters c and γ of the RBF kernel using a 5-fold cross-validation on the training set. The thresholds r_1, $r_{2,1}$, and $r_{2,2}$ used to partition the label space (i.e., the age range) are empirically determined.

2.4 Face Image Quality Assessment

It is well known that the performance of a face recognition system depends on the quality of face images. While face image quality is not easy to define, it is influenced by factors such as IPD, pose, blur, illumination, etc. [69], [70]. We also notice this to be true for demographic estimation. In this section, we present a learning-based quality assessment (QA) method to identify and reject low-quality face images. Quality assessment is mainly to detect low-quality face images due to variations of pose, illumination, occlusion, blur, etc. However, these low-quality images may also include visually satisfactory face images

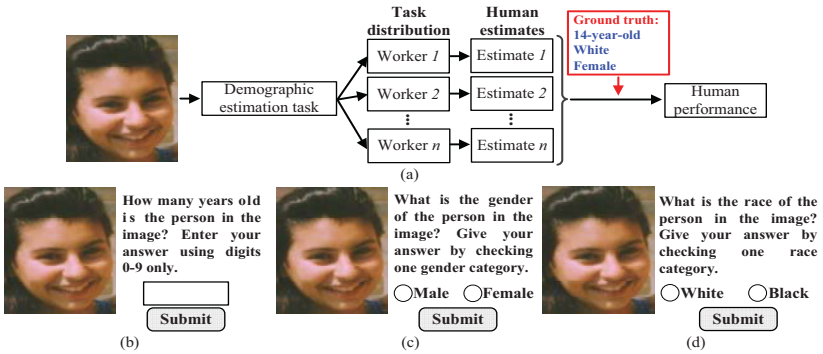


Fig. 8. Demographic estimation obtained using the MTurk crowdsourcing service: (a) overview of the estimation process, and illustrations of three HITs for (b) age estimation, (c) gender classification, and (d) race classification. Images shown to the MTurk workers are exactly the same as those were input to our algorithm.

that the system is less likely to make estimates with high confidence (e.g., samples near the age group boundaries).

The training set for learning the hierarchical demographic estimator in Section 2.3 is also used to build our quality assessment model. Taking age estimation as an example, we partition the training set into high-quality (positive) and low-quality (negative) subsets based on a threshold P for the training set age estimation error (see Appendix B). However, it is inevitable that the positive and negative sets are severely imbalanced. To address this problem, resampling with replacement is used for the positive samples. In each resampling of positive samples, the same number of positive samples is drawn as the negative samples. This resampling is performed (K) times to build an ensemble classifier to distinguish between high-quality and low-quality face images,

$$QA(\mathbf{x}) = \frac{1}{K}\sum_{k=1}^{K} Q_k(\mathbf{x}), \tag{7}$$

where $Q_k(\cdot)$ is a binary SVM classifier with the RBF kernel, and $Q_k(\cdot) = 1$ if a face image is of high quality, and $Q_k(\cdot) = 0$, otherwise. Based on the ensemble classifier, we reject a test face image $\mathbf{x}_t$ only when $QA(\mathbf{x}_t) = 0$. Such a conservative rejection criterion assures a low rejection rate during testing but is still effective in improving the overall accuracy of demographic estimation. In our age estimation experiments, $K = 5$ is used, and P is set so that we reject about 5% of the test images. The proposed quality assessment approach is also applied on the datasets used in human age estimation experiments, for a fair comparison of human vs. machine performance. Given the high accuracy of our method for gender and race classification, no quality assessment is done for these two problems.

3 Demographic Estimation by Human

3.1 Design of Human Estimation Tasks

As a baseline, we gather demographic estimates made by human workers using the Amazon Mechanical Turk (MTurk) crowdsourcing service[4]. The human intelligence tasks (HITs) consist of displaying a face image with a prompt string asking a single question about the demographics of the person in the face image (see Fig. 8). The GUI was designed to constrain user's input to a valid age range for age estimation, and to binary choices for gender and race classification; this cuts down unintentional mistakes by the MTurk workers. For each task, we enforced that each MTurk worker could provide only a single response to avoid bias. All the face images were stored in our lab's server; Adobe SWF is used to display face images to the MTurk workers. This way, the face images could not be downloaded by MTurk workers.

We collected age estimates provided by the crowd for the entire FG-NET database $(1,002$ images), a subset of MORPH II $(2,000$ images), and a subset of PCSO $(2,200$ images) with 10 MTurk workers per image. There is no way to measure the human gender and race estimation accuracies on FG-NET because the real gender and race of subjects in FG-NET are not available. Therefore, gender and race estimations by the crowd are performed only on subsets of MORPH II and PCSO databases, with three workers per image. For race classification on PCSO, we use a different $2,000$-image subset to ensure a balance of white and black subjects (see Appendix A). Age, gender and race estimates were also collected for a subset of LFW $(4,211$ images) with three workers per image. In total, we posted $112,519$ HITs on MTurk. The payment for each age estimation HIT was 3 cents, and the payment for each gender and race classification HIT was 2 cents for a total cost of about $\$3,000$.

3.2 Crowdsourced Response Data

As is typical of crowdsourcing experiments, some of the response data is noisy and unusable. For example, in age estimation without constraining user's input, some workers submitted an empty text box, a 3-digit age, an age range instead of a specific age (e.g., 52–60), or age in words such as "forty". When workers

4. https://www.mturk.com

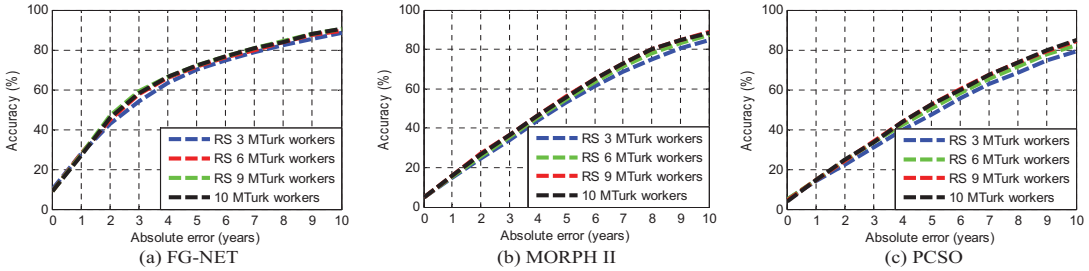


Fig. 9. The difference in human age estimation performance between using 3, 6, and 9 randomly selected (RS) MTurk workers and all 10 workers.

entered an age range, we used the middle of the range as their estimate, and we manually converted the word entries to integers. We rejected the remaining problematic cases (e.g., empty text boxes, age over 100), and obtained replacement estimates from additional workers. Out of the initial 10,020 FG-NET age estimates, we found only 16 to be unusable, and out of the initial 22,000 PCSO age estimates, we found only 604 to be unusable[5].

Given the 10 age estimates of each face image, a straightforward way to compute the error of the human estimates would be to discard the highest and lowest age estimates for each image, take the mean of the remaining 8 estimates, and compute the mean absolute error (MAE) [15]. However, such a strategy does not represent the performance of individual human workers; rather it represents the performance of a group of workers collectively. Therefore, in our experiments, we directly calculate the MAE of 8 individual age estimates (still discarding the highest and lowest estimates to reduce the impact of input errors by workers) w.r.t. the ground-truth age.

One question regarding crowdsourcing is how many estimates per image are sufficient. Since we initially collected 10 human age estimates per image (for FG-NET, a subset of MORPH II, and a subset of PCSO), we randomly select 3, 6, and 9 human age estimates for each image, and calculate the average age for each image. We repeat each experiment five times, report the average age estimates with 3, 6, and 9 MTurk workers, and compare them with those using all 10 MTurk workers. The results in Fig. 9 show that three MTurk workers per image are sufficient for the age estimation task.

For gender and race classifications, a worker has less room to make errors than the age estimation. We believe three MTurk workers per image are also sufficient for the human experiments on gender and race classification tasks. So for gender and race classification results, we report the majority choice among the three HITs per image, which is a reasonable binary decision rule based on May's theorem [72].

5. The human age estimates of face images in the FG-NET database are available at: http://www.cse.msu.edu/rgroups/biometrics/pubs/databases.html

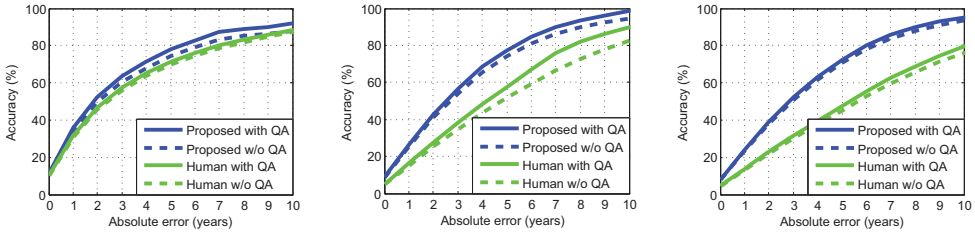
4 Experimental Results

Our approach for demographic estimation tasks has been evaluated on the FG-NET [20], FERET [38], MORPH Album2 [63], PCSO, and LFW [71] databases summarized in Appendix A. Human demographic estimates obtained via crowdsourcing are also reported as baselines, which facilitate an understanding of the differences between human (crowdsourced worker) and our algorithm in demographic estimation from a face image. Additionally, our approach is also compared with the state of the art (Tables 1–3).

4.1 Age Estimation

Table 4 lists the Mean Absolute Error (MAE) of age estimation by our approach and human. Without quality assessment (QA) applied to FG-NET, the MAE of human is 4.7 years, which is slightly better than that of the proposed approach (4.8 years). However, with QA applied to FG-NET, our system achieves much lower MAE (3.8 years) than human (4.5 years). While the MAEs of human estimates on MORPH II and PCSO without QA are 6.3 and 7.2 years, respectively, the MAEs of the proposed approach without QA are 3.8 and 4.3 years, respectively. With QA to reject some low-quality images (see Fig. 4 in Appendix B), the MAEs are further reduced for both human (4.3 and 6.6 years) and the proposed approach (3.6 and 4.1 years). On MORPH II and PCSO, the proposed age estimation approach always performs better than human. In Fig. 11, we show the correlations between the ground-truth ages (true ages), and age estimates by the proposed approach and human on the MORPH II subset with 2,000 images. We notice that while human tends to overestimate the ages of individual subjects (Fig. 11 (b)), the proposed approach provides relatively unbiased estimates (Fig. 11 (a)).

The main reason why human outperforms the proposed approach (in terms of MAE) on FG-NET without QA is the significantly biased age distribution (a majority of images belong to subjects less than 18 years of age) in FG-NET (see Fig. 1 (a) in Appendix A). As a result, the training set used to train our algorithm does not contain a sufficient number of images in the age range 30–69 (∼4.5 images per age). In contrast, both MORPH II and PCSO have subjects with more



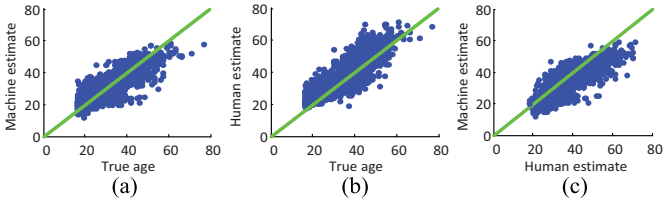
(a) FG-NET (b) MORPH II (c) PCSO

Fig. 10. Age estimation performance with and without quality assessment by the proposed approach and human.

TABLE 4
Mean Absolute Error (MAE) of age estimation on FG-NET, MORPH II, and PCSO databases (in years).

Database	Proposed algorithm		Human workers	
	w/o QA	w/ QA	w/o QA	w/ QA
FG-NET	4.8±6.2	3.8±4.2	4.7±5.0	4.5±4.8
MORPH II	3.8±3.3	3.6±3.0	6.3±4.9	4.3±3.8
PCSO	4.3±3.7	4.1±3.3	7.2±5.7	6.6±4.9

We reject 5% of all the test images when quality assessment (QA) is used. For both the proposed algorithm and human workers, standard deviation is calculated from all the face images, because MTurk workers cannot do demographic estimation following a 5-fold cross-validation protocol.



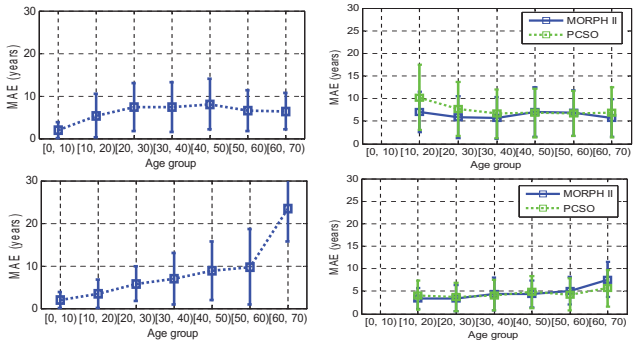
(a) (b) (c)

Fig. 11. Correlations between (a) age estimates by the proposed algorithm vs. true ages, (b) age estimates by human vs. true ages, and (c) age estimates by the proposed algorithm vs. human on the MORPH II subset with 2,000 images.

uniform age distributions. Hence, our approach has consistently low MAEs across individual age ranges (see Fig. 12 (b)). These comparisons suggest that the age estimation performance of human is relatively independent of the age distribution in the database and the database size because of their prior knowledge.

While MAE reflects the overall performance of an age estimation method, it does not explicitly reveal the system accuracy within a particular age error range, which is important for practical applications. Therefore, in Fig. 10 we show the Cumulative Score (CS) [19] of age estimation by human and the proposed approach within 0–10 years absolute error. Our approach performs consistently better than human on all the three databases with and without QA.

The first row in Fig. 15 shows examples of good and poor age estimates by our approach and human. The first row of Fig. 15 (d) reveals an interesting phenomenon: the estimates from the proposed method are poor compared with the ground-truth ages but are fairly consistent with human perception. In these examples, one may question the accuracy of the labeled ages in the database. For example, Fig. 5 in Appendix



(a) FG-NET (b) MORPH II & PCSO

Fig. 12. Per age range MAE of age estimation by human (top) and the proposed algorithm (bottom).

C shows two images where human estimation errors are large (≥ 30 years), but obviously the labeled (real) ages in the database are wrong.

Table 1 lists age estimation results from state-of-the-art methods on FG-NET and MORPH II, where MAE and the "CS at 5-year absolute error" are reported. The best known MAE performance on FG-NET is 4.1 years [27], which is better than our approach without quality assessment (MAE of 4.8 years). However, as show in Fig. 14, our approach performs much better than [27] in the useful (operational) error range of 0–5 years absolute error[6]. A comparison with [27] on MORPH II cannot be done, because MORPH II was not evaluated in [27]. The best known performance on MORPH II as reported in the literature is 4.0-year MAE [33], which is the same as our approach without quality assessment[7], but worse than our method with quality assessment (MAE of 3.7 years). Compared to [33], we use a larger set of MORPH II (78,207 images vs. 55,132 images in [33]). Additionally, while the training set in [33] is constructed with relatively balanced race and gender groups, the training set in our experiments is randomly selected from the MORPH II database. The CS curve on MORPH II was not provided in [33], so we cannot compare the two methods in the operational error range of 0–5 years.

In the above experiments, we always use the real age, also called *chronological age*, as the ground-truth

6. An absolute error larger than five years defines an age range larger than 10 years, which is less helpful in practical applications.

7. To compare with [33], we reduced the size of our training set by randomly selecting 20% of the training images.

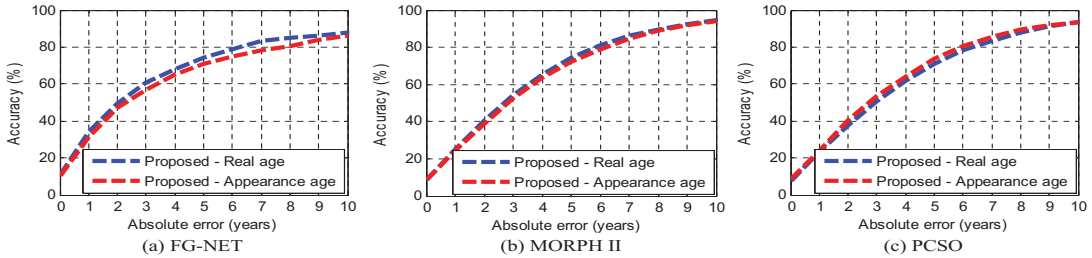


Fig. 13. Performance of the proposed age estimation method with appearance (human estimated) age (red curve) and real age (blue curve) as the ground-truth age. No quality assessment is applied in these experiments.

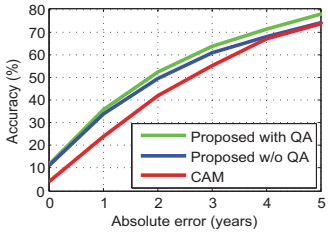


Fig. 14. A comparison of the proposed approach and the state-of-the-art method, CAM [27] on FG-NET.

age. If the real age is not available, the *appearance age*, which is the age perceived by human, has to be used as the ground truth to train the age estimation system. We have also evaluated our age estimation algorithm by using appearance age as the ground-truth age for (i) the FG-NET database, (ii) a subset from MORPH II (2,000 images), and (iii) a subset from PCSO (2,200 images). Figure 13 shows that the age estimates by our method are fairly robust to whether appearance age or real age is used as the ground truth. One reason for this is that appearance age is highly correlated with the real age (see Fig. 11 (b)).

4.2 Gender Classification

In this section, the proposed approach is evaluated on gender classification, and compared with the human performance. Since no gender information is provided with images in FG-NET, gender classification experiments are performed on MORPH II and PCSO. The FERET database is also used to compare the proposed approach with published methods. We do not perform human estimation of gender on FERET due to its smaller size than other public-domain databases, such as MORPH II.

The gender classification results obtained from the proposed approach and human workers are shown in Table 5. The table includes the overall accuracy (males vs. females), and the 2×2 confusion matrix for each method. Based on the overall accuracy, our approach performs slightly better than human on both MORPH II (97.6% vs. 96.9%) and PCSO (97.1% vs. 96.5%) databases. Regarding the per gender accuracy, the misclassification rate of females is higher than that of males for both our approach and human. For example, on MORPH II, the misclassification rates

TABLE 5
Confusion matrix for gender classification (in %).

Database	Proposed method		True gender	Human workers	
	Female	Male		Female	Male
MORPH II	91.9	8.1	Female	94.1	5.9
	1.3	98.7	Male	0.4	99.6
Overall	97.6			96.9	
PCSO	95.7	4.3	Female	93.5	6.5
	2.4	97.6	Male	0.4	99.6
Overall	97.1			96.5	
FERET	94.8	5.2	Female	N/A	
	2.0	98.0	Male		
Overall	96.8			N/A	

for females by our approach and human are 8.1% and 5.9%, respectively, which are much higher than the misclassification rates for males, 1.3% and 0.4%, respectively. This is not surprising because many important facial features of females, such as the hairstyle and the shape of eyebrows, may change frequently. The second row in Fig. 15 gives some examples of correct and incorrect gender classification results.

Table 2 lists the gender classification accuracies of state-of-the-art methods. The best known accuracy on the MORPH II database is 96.0% [33]. Our approach achieves a higher accuracy of 97.6%, which is a 40% reduction in the gender classification error reported in [33]. Additionally, while a subset of 55,132 face images was used in [33], we use a larger MORPH II subset containing 78,207 images. The best known accuracy for gender classification on FERET is 99.1% [54], which is higher than our accuracy of 96.8%. However, we should point out that only 199 female and 212 male images were used in [54], while we use a larger FERET subset with 1,007 female and 1,712 male images.

4.3 Race Classification

We conducted experiments on MORPH II and PCSO databases to distinguish between black and white subjects, who constitute the majority of subjects in these two databases. Results in Table 6 show that our approach consistently outperforms human on both databases. For example, the proposed approach achieves 1.3% (99.1% vs. 97.8%) and 2.2% (98.7% vs. 96.5%) higher overall accuracies than human on

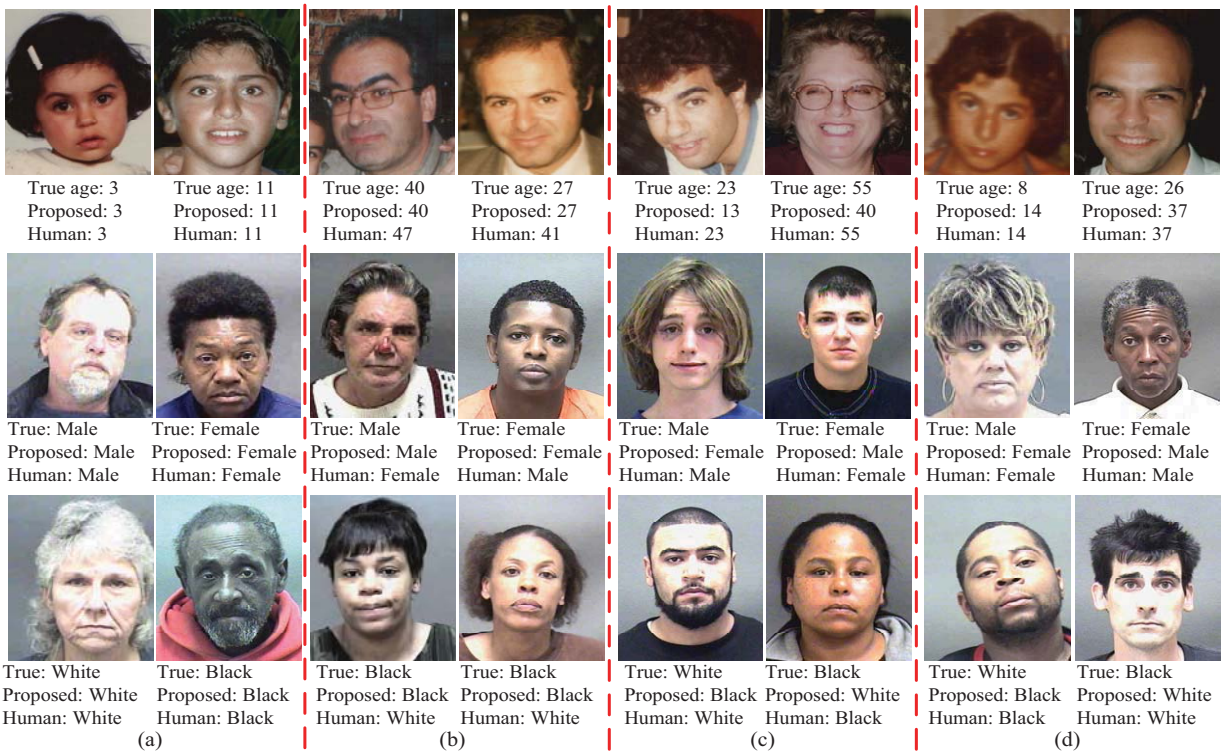
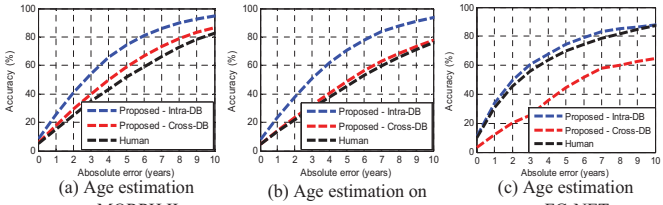


Fig. 15. Examples of good and poor demographic estimates for age (top row), gender (middle row), and race (bottom row). (a) Proposed algorithm and human workers provide good estimates; (b) Proposed algorithm provides good estimates but human workers do not; (c) Proposed algorithm provides poor estimates compared to human workers; and (d) Both the proposed algorithm and human workers provide poor estimates.

TABLE 6
Confusion matrix for race classification (in %).

Database	Proposed method		True Race	Human workers	
	Black	White		Black	White
MORPH II	99.1	0.9	Black	95.9	4.1
	1.1	98.9	White	0.3	99.7
Overall	99.1			97.8	
PCSO	97.5	2.5	Black	97.3	2.7
	0.9	99.1	White	0.8	99.2
Overall	98.7			96.5	



(a) Age estimation on MORPH II (b) Age estimation on PCSO (c) Age estimation on FG-NET

Testing database	Classification task	Proposed - Intra-DB	Proposed - Cross-DB	Human
MORPH II	Gender	97.6	90.5	96.9
	Race	99.1	95.3	97.8
PCSO	Gender	97.1	88.6	96.5
	Race	98.7	93.0	96.5

(d) Gender and race classification on MORPH II and PCSO (in %)

Fig. 16. Cross-database testing on MORPH II, PCSO and FG-NET for three demographic estimation tasks. No quality assessment is applied in these experiments.
* The training sets for the cross-database testing on MORPH II, PCSO and FG-NET are PCSO, MORPH II, and MORPH II and PCSO, respectively.

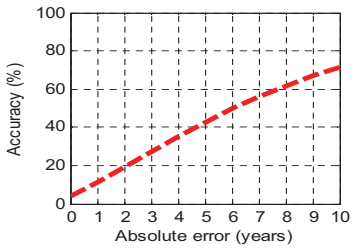
MORPH II and PCSO, respectively. While our approach performs better than human in classifying black subjects, human is relatively better at classifying white subjects. The third row in Fig. 15 gives some examples of correct and incorrect race classifications by our approach and human.

Race classification accuracies from state-of-the-art methods are shown in Table 3. The best known accuracy for white vs. black classification on MORPH II database (with $55,132$ images) is 98.9% [33]. Our approach achieves a higher accuracy of 99.1% on a larger MORPH II subset with $78,207$ images.

4.4 Generalization Ability

We evaluate the generalization ability of the proposed approach using: (i) cross-database testing on the MORPH II, PCSO, and FG-NET databases; and (ii) demographic estimation on a subset ($4,211$ images) of the more challenging LFW database [71].

The demographic estimation results with cross-database testing on MORPH II, PCSO, and FG-NET are shown in Fig. 16. As expected, cross-database testing performance is lower than intra-database testing. But, we believe these accuracies (not reported in other published studies) are still quite good. Image quality (resolution and illumination) differences between the PCSO and MORPH databases are responsible for the drop in performance. This suggests the need for training on a larger representative database encompassing



(a) Age estimation

Overall 94.4	Female	Male
Female	89.5	10.5
Male	3.8	96.2

(b) Gender classification (in %)

Overall 90.8	White	Other
White	91.5	8.5
Other	12.4	87.6

(c) Race classification (in %)

Fig. 17. Performance of the proposed demographic estimation approach on a subset of the LFW database by using human (crowdsourced) estimates of age, gender, and race as the ground truth. No quality assessment is applied in these experiments.

more variety in imaging conditions if the system is applied in arbitrary settings.

It is reasonable to compare the cross-database performance of the proposed approach with human performance, since human performance is not tuned to a particular database. For age estimation, the proposed approach with cross-database testing performs slightly better than human on MORPH II and PCSO, but much worse on FG-NET. For gender and race classification, human outperforms the proposed approach with cross-database testing. This is due to (i) the significantly different age distributions between FG-NET and the training dataset (MORPH II and PCSO), and (ii) the prior knowledge of human.

The LFW database was collected for studying the problem of unconstrained face recognition. We provide demographic estimation results on a subset ($4,211$ images) of the LFW database, where the face images have relatively small pose variations (see Fig. 2 in Appendix A). The human estimates of age, gender, and race (see the demographic distributions in Appendix A) are used as the ground truth to evaluate the proposed approach. Demographic estimation results are shown in Fig. 17. A baseline performance was reported by Kumar *et al.* [2], where age group, gender, and race were utilized as three of the 73 attributes to distinguish different subjects. The sizes of the training set in our approach and that in [2] are comparable, and SVM classifiers were also used in [2]. Our gender classification method achieves better performance (94%) than that reported in [2] (91.38%), which shows the importance of feature representation. These results show that the proposed approach generalizes very well to the challenging LFW face database.

4.5 Prototype System

To enable real-time demographic estimation (e.g., using a webcam), a prototype system of our approach is implemented using C++. Online testing using the prototype system involves the following steps: 1) face detection and preprocessing (FP); 2) demographic informative feature extraction (DIFE); 3) quality assessment (QA); and 4) hierarchical estimation, including group classification (GC) and within-group regression (GR). The computational cost of each step is given in Table 7. Offline training takes ~ 10 hours using a training set of $10K$ images from MORPH II. Our system operates at about 10 fps for $720p$ videos; this speed on a commodity laptop is achieved primarily due to the proposed feature selection and hierarchical estimation methods.

TABLE 7
Average time (sec.) per algorithmic step in our prototype demographic estimation system.

FP	DIFE	QA	GC		GR	**Total**
			B_1	$B_{2,*}$	$R_{*,*}$	
0.02	0.03	0.005	0.025	0.005	0.005	$\sim$0.09

A laptop with 2.9 GHz dual-core Intel Core i7 processor and 8G RAM was used.

Among the state-of-the-art methods for joint age, gender, and race estimation [7], [8], [33], only [33] reports the computational cost without face detection and feature extraction (1.6 sec. per image), which is 40 times slower than the corresponding components of the proposed approach (0.04 sec. per image).

5 Conclusions

This paper presents a generic framework for automatic demographic (age, gender and race) estimation from a given face image. We extract demographic informative features from the commonly used biologically inspired features (BIF), and predict the demographic attributes of a face image using a hierarchical classifier. Quality assessment is proposed to identify low-quality face images, which allows possible reacquisition of new face images in cooperative scenarios, or rejection of the input face image otherwise. Human ability to estimate age, gender and race from the same face images that are processed by our algorithm is also evaluated using crowdsourced data obtained via the Amazon Mechanical Turk (MTurk) service. A comparison shows that our algorithm can closely match human performance in demographic estimation.

Our approach performs well on large and diverse databases (including FG-NET, FERET, MORPH II, PCSO, and LFW), and performs slightly better than human on MORPH II and PCSO in all the three demographic estimation tasks. Our approach also outperforms state-of-the-art methods in most of the experiments reported here. A prototype system of our demographic estimation algorithm (C++ implementation) illustrates the feasibility of performing age, gender and race estimation in real time using a commodity processor. Future work includes the study of the other-race effect in human perception of demographics, and automatic demographic estimation from unconstrained face images.

ACKNOWLEDGMENTS

The authors would like to thank the Pinellas County Sheriff's Office for providing the PCSO database, and Dr. Guodong Guo for providing the BIF features of the FG-NET database as a reference. This manuscript benefited from the valuable comments provided by the reviewers. All correspondence should be directed to Anil K. Jain.

REFERENCES

[1] S.Z. Li and A.K. Jain (eds.), *Handbook of Face Recognition*, 2nd ed. New York: Springer, 2011.
[2] N. Kumar, A. Berg, P. Belhumeur, and S. Nayar, "Describable Visual Attributes for Face Verification and Image Search," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 10, pp. 1962-1977, Oct. 2011.
[3] P.J. Phillips and A.J. O'Toole, "Comparison of Human and Computer Performance across Face Recognition Experiments," *Image Vision Comput.*, vol. 32, no. 1, pp. 74-85, Jan. 2014.
[4] L. Best-Rowden, S. Bisht, J. Klontz and A. K. Jain, "Unconstrained Face Recognition: Establishing Baseline Human Performance via Crowdsourcing," *Proc. IJCB*, Sept. 2014.
[5] W.J. Scheirer, N. Kumar, P.N. Belhumeur, and T.E. Boult, "Multi-Attribute Spaces: Calibration for Attribute Fusion and Similarity Search," *Proc. IEEE CVPR*, pp. 2933-2940, 2012.
[6] N. Chhaya and T. Oates, "Joint Inference of Soft Biometric Features", *Proc. ICB*, pp. 466-471, 2012.
[7] Z. Yang and H. Ai, "Demographic Classification with Local Binary Patterns," *Proc. ICB*, pp. 464-473, 2007.
[8] A. Hadid and M. Pietikänen, "Demographic Classification from Face Videos Using Manifold Learning," *Neurocomputing*, vol. 100, pp. 197-205, Jan. 2013.
[9] L.T. Semaj, "The Development of Racial-Classification Abilities," *J. Negro Ed.*, vol. 50, no. 1, pp. 41-47, Winter, 1981.
[10] D.M. Burt and D.I. Perrett, "Perception of Age in Adult Caucasian Male Faces: Computer Graphic Manipulation of Shape and Colour Information," *Proc. Biol Sci.*, vol. 259, no. 1355, pp. 137-143, Feb. 1995.
[11] M. Rhodes, "Age Estimation of Faces: A Review," *Appl. Cognit. Psychol.*, vol. 23, no. 1, pp. 38-59, Jan. 2009.
[12] G.W. Cottrell and J. Metcalfe, "EMPATH: Face, Emotion, and Gender Recognition Using Holons," *Proc. NIPS*, pp. 564-571, 1990.
[13] Y.H. Kwon and N.V. Lobo, "Age Classification from Facial Images," *Comput. Vis. Image Und.*, vol. 74, no. 1, pp. 1-21, Apr. 1999.
[14] N. Ramanathan and R. Chellappa, "Modeling Age Progression in Young Faces," *Proc. IEEE CVPR*, pp. 387-394, 2006.
[15] M.C. Voelkle, N.C. Ebner, U. Lindenberger, and M. Riediger, "Let Me Guess How Old You Are: Effects of Age, Gender, and Facial Expression on Perceptions of Age," *Psychol. Aging.*, vol. 27, no. 2, pp. 265-277, Jun. 2012.
[16] Y. Fu, G. Guo, and T.S. Huang, "Age Synthesis and Estimation via Faces: A Survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 11, pp. 1955-1976, Nov. 2010.
[17] A. Lanitis, C. Taylor, and T. Cootes, "Toward Automatic Simulation of Aging Effects on Face Images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 12, no. 4, pp. 442-455, Apr. 2002.
[18] J. Hayashi, M. Yasumoto, H. Ito, and H. Koshimizu, "Age and Gender Estimation Based on Wrinkle Texture and Color of Facial Images" *Proc. ICPR*, pp. 405-408, 2002.
[19] X. Geng, Z.-H. Zhou, and K. Smith-Miles, "Automatic Age Estimation Based on Facial Aging Patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 12, pp. 2234-2240, Dec. 2007.
[20] Facial Image Processing and Analysis (FIPA), "FG-NET Aging Database," http://fipa.cs.kit.edu/433.php#Downloads.
[21] Y. Fu and T.S. Huang, "Human Age Estimation with Regression on Discriminative Aging Manifold," *IEEE Trans. Multimedia*, vol. 10, no. 4, pp. 578-584, Jun. 2008.
[22] J. Suo, T. Wu, S.-C. Zhu, S. Shan, X. Chen, and W. Gao, "Design Sparse Features for Age Estimation Using Hierarchical Face Model," *Proc. FGR*, pp. 1-6, 2008.
[23] G. Guo, G. Mu, Y. Fu, and T.S. Huang, "Human Age Estimation Using Bio-inspired Features," *Proc. IEEE CVPR*, pp. 112-119, 2009.
[24] M. Riesenhuber and T. Poggio, "Hierarchical Models of Object Recognition in Cortex," *Nat. Neurosci.*, vol. 2, no. 11, pp. 1019-1025, Nov. 1999.
[25] G. Guo and G. Mu, "Simultaneous Dimensionality Reduction and Human Age Estimation via Kernel Partial Least Squares Regression," *Proc. IEEE CVPR*, pp. 657-664, 2011.
[26] S.E. Choi, Y.J. Lee, S.J. Lee, K.R. Park, and J. Kim, "Age Estimation Using a Hierarchical Classifier Based on Global and Local Facial Features," *Pattern Recogn.*, vol. 44, no. 6, pp. 1262-1281, Jun. 2011.
[27] K. Luu, K. Seshadri, M. Savvides, T. Bui, and C. Suen, "Contourlet Appearance Model for Facial Age Estimation," *Proc. IJCB*, pp. 1-8, 2011.
[28] K.-Y. Chang, C.-S. Chen, and Y.-P. Hung. "Ordinal Hyperplanes Ranker with Cost Sensitivities for Age Estimation," *Proc. IEEE CVPR*, pp. 585-592, 2011.
[29] T. Wu, P. Turaga, and R. Chellappa, "Age Estimation and Face Verification across Aging Using Landmarks," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 6, pp. 1780-1788, Dec. 2012.
[30] P. Thukral, K. Mitra, and R. Chellappa, "A Hierarchical Approach for Human Age Estimation," *Proc. IEEE ICASSP*, pp. 1529-1532, 2012.
[31] J. Lu and Y. Tan, "Ordinary Preserving Manifold Analysis for Human Age and Head Pose Estimation," *IEEE Trans. Human-Mach. Syst.*, vol. 43, no. 2, pp. 249-258, Mar. 2013.
[32] W.-L. Chao, J.-Z. Liu, and J.-J. Ding, "Facial Age Estimation Based on Label-sensitive Learning and Age-oriented Regression," *Pattern Recogn.*, vol. 46, no. 3, pp. 628-641, Mar. 2013.
[33] G. Guo and G. Mu, "Joint Estimation of Age, Gender and Ethnicity: CCA vs. PLS," *Proc. FGR*, pp. 1-6, 2013.
[34] X. Geng, C. Yin, and Z.-H. Zhou, "Facial Age Estimation by Learning from Label Distributions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 10, pp. 2401-2412, Oct. 2013.
[35] E. Mäkinen and R. Raisamo, "An Experimental Comparison of Gender Classification Methods," *Pattern Recogn. Lett.*, vol. 29, no. 10, pp. 1544-1556, Jul. 2008.
[36] R. Brunelli and T. Poggio, "HyberBF Networks for Gender Classification," *Proc. DARPA Image Understanding Workshop*, pp. 311-314, 1995.
[37] B. Moghaddam and M. Yang, "Gender Classification with Support Vector Machines," *Proc. FGR*, pp. 306-311, 2000.
[38] P.J. Phillips, H. Moon, S. A. Rizvi, P.J. Rauss, "The FERET Evaluation Methodology for Face-Recognition Algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 10, pp. 1090-1104, Oct. 2000.
[39] S. Gutta, J.J. Huang, P. Jonathon, and H. Wechsler, "Mixture of Experts for Classification of Gender, Ethnic Origin, and Pose of Human Faces," *IEEE Trans. Neural Netw.*, vol. 11, no. 4, pp. 948-960, Jul. 2000.
[40] P. Viola and M.J. Jones, "Robust Real-time Object Detection," *Proc. ICCV*, pp. 1254-1259, 2001.
[41] B. Wu, H. Ai, and C. Huang, "LUT-based AdaBoost for Gender Classification," *Proc. AVBPA*, pp. 104-110, 2003.
[42] T.F. Cootes, G.J. Edwards, and C.J. Taylor, "Active Appearance Models," *Proc. ECCV*, pp. 484-498, 1998.
[43] Y. Saatci and C. Town, "Cascaded Classification of Gender and Facial Expression Using Active Appearance Models," *Proc. FGR*, pp. 393-400, 2006.
[44] S. Baluja and H.A. Rowley, "Boosting Sex Identification Performance," *Int. J. Comput. Vision*, vol. 71, no. 1, pp. 111-119, Jan. 2007.
[45] W. Gao and H. Ai, "Face Gender Classification on Consumer Images in a Multiethnic Environment," *Proc. ICB*, pp. 169-178, 2009.
[46] J. Wu, W.A.P. Smith, and E.R. Hancock, "Facial Gender Classification Using Shape-from-shading," *Comput. Vis. Image Und.*, vol. 28, no. 6, pp. 1039-1048, Jun. 2010.
[47] S. Mozaffari, H. Behravan, and R. Akbari, "Gender Classification using Single Frontal Image per Person: Combination of Appearance and Geometric Based Features," *Proc. ICPR*, pp. 1192-1195, 2010.

[48] Y. Wang, K. Ricanek, C. Chen, and Y. Chang, "Gender Classification from Infants to Seniors," *Proc. BTAS*, pp. 1-6, 2010.

[49] Y. Dong and D.L. Woodard, "Eyebrow Shape-Based Features for Biometric Recognition and Gender Classification: A Feasibility Study," *Proc. IJCB*, pp. 1-8, 2011.

[50] G. Zhang and Y. Wang, "Hierarchical and Discriminative Bag of Features for Face Profile and Ear based Gender Classification," *Proc. IJCB*, pp. 1-8, 2011.

[51] J. Bekios-Calfa, J.M. Buenaposada, and L. Baumela, "Revisiting Linear Discriminant Techniques in Gender Recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 4, pp. 858-864, Apr. 2011.

[52] L. Ballihi, B.B. Amor, M. Daoudi, A. Srivastava, and D. Aboutajdine, "Boosting 3-D-Geometric Features for Efficient Face Recognition and Gender Classification," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 6, pp. 1766-1779, Dec. 2012.

[53] C. Chen and A. Ross, "Local Gradient Gabor Pattern (LGGP) with Applications in Face Recognition, Cross-spectral Matching and Soft Biometrics," *Proc. SPIE*, May 2013.

[54] J.E. Tapia and C.A. Perez, "Gender Classification Based on Fusion of Different Spatial Scale Features Selected by Mutual Information from Histogram of LBP, Intensity, and Shape," *IEEE Trans. Inf. Forensics Security*, vol. 8, no. 3, pp. 488-499, Mar. 2013.

[55] U. Tariq, Y. Hu, and T.S. Huang, "Gender and Ethnicity Identification from Silhouetted Face Profiles," *Proc. ICIP*, pp. 2413-2416, 2009.

[56] H. Han, C. Otto, and A.K. Jain, "Age Estimation from Face Images: Human vs. Machine Performance," *Proc. ICB*, pp. 1-8, 2013.

[57] Cognitec Systems GmbH, "FaceVACS Software Developer Kit," http://www.cognitec-systems.de, 2010.

[58] T. Serre, L. Wolf, and T. Poggio, "Object Recognition with Features Inspired by Visual Cortex," *Proc. IEEE CVPR*, pp. 994-1000, 2005.

[59] J. Mutch and D. Lowe, "Object Class Recognition and Localization Using Sparse Features with Limited Receptive Fields," *Proc. IEEE CVPR*, pp. 45-57, 2006.

[60] E. Meyers and L. Wolf, "Using Biologically Inspired Features for Face Processing," *Int. J. Comput. Vis.*, vol. 76, no. 1, pp. 93-104, Jan. 2008.

[61] C. Liu and H. Wechsler, "Gabor Feature based Classification Using the Enhanced Fisher Linear Discriminant Model for Face Recognition," *IEEE Trans. Image Process.*, vol. 11, no. 4, pp. 467-476, Apr. 2002.

[62] V.N. Vapnik, *Statistical Learning Theory*, New York: John Wiley, 1998.

[63] K. Ricanek and T. Tesafaye, "MORPH: A Longitudinal Image Database of Normal Adult Age-progression," *Proc. FGR*, pp. 341-345, 2006.

[64] D.H. Hubel and T.N. Wiesel, "Receptive Fields, Binocular Interaction and Functional Architecture in the Cat's Visual Cortex," *J. Physiol.*, vol. 160, no. 1, pp. 106-154, Jan. 1962.

[65] Y. Freund and R.E. Schapire, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," *J. Comput. Syst. Sci.*, vol. 55, no. 1, pp. 119-139, Aug. 1997.

[66] K. Tsukahara, K. Sugata, O. Osanai, A. Ohuchi, Y. Miyauchi, M. Takizawa, M. Hotta, and T. Kitahara, "Comparison of Age-related Changes in Facial Wrinkles and Sagging in the Skin of Japanese, Chinese and Thai Women," *J. Dermatol. Sci.*, vol. 47, no. 1, pp. 19-28, Jul. 2007.

[67] E. Brown and D.I. Perrett, "What Gives a Face Its Gender?" *Perception*, vol. 22, no. 7, pp. 829-840, Jul. 1993.

[68] L. Zhao and S. Bentin, "The Role of Features and Configural Processing in Face-race Classification," *Vision Res.*, vol. 51, no. 23-24, pp. 2462-2470, Dec. 2011.

[69] E. Tabassi, C. Wilson, and C. Watson, "Fingerprint Image Quality," *NIST Research Report NISTIR7151*, Aug. 2004.

[70] P. Grother and E. Tabassi, "Performance of Biometric Quality Measures," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 4, pp. 531-543, Apr. 2007.

[71] G.B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments," *University of Massachusetts, Amherst, Technical Report 07-49*, Oct. 2007.

[72] K. O. May, "A Set of Independent Necessary and Sufficient Conditions for Simple Majority Decisions," *Econometrica*, vol. 20, no. 4, pp. 680-684, Oct. 1952.

Hu Han is a Research Associate in the Department of Computer Science and Engineering at Michigan State University, East Lansing. He received the B.S. degree from Shandong University, Jinan, China, and the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2005 and 2011, respectively, both in Computer Science. His research interests include computer vision, pattern recognition, and image processing, with applications to biometrics, forensics, law enforcement, and security systems. He is a member of the IEEE.

Charles Otto received his B.S. degree in the Department of Computer Science and Engineering at Michigan State University in 2008. He was a research engineer at IBM in 2006-2011. Since 2012, he has been working toward the Ph.D. degree in the Department of Computer Science and Engineering at Michigan State University. His research interests include pattern recognition, image processing, and computer vision, with applications to face recognition. He is a student member of the IEEE.

Xiaoming Liu is an Assistant Professor in the Department of Computer Science and Engineering at Michigan State University (MSU). He received the B.E. degree from Beijing Information Technology Institute, China and the M.E. degree from Zhejiang University, China, in 1997 and 2000 respectively, both in Computer Science, and the Ph.D. degree in Electrical and Computer Engineering from Carnegie Mellon University in 2004. Before joining MSU in Fall 2012, he was a research scientist at General Electric Global Research Center. His research areas are face recognition, biometrics, image alignment, video surveillance, computer vision and pattern recognition. He has authored more than 70 scientific publications, and has filed 22 U.S. patents. He is a member of the IEEE.

Anil K. Jain is a University Distinguished Professor in the Department of Computer Science and Engineering at Michigan State University, East Lansing. His research interests include pattern recognition and biometric authentication. He served as the editor-in-chief of the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE (1991-1994). He is the coauthor of a number of books, including Handbook of Fingerprint Recognition (2009), Handbook of Biometrics (2007), Handbook of Multibiometrics (2006), Handbook of Face Recognition (2005), BIOMETRICS: Personal Identification in Networked Society (1999), and Algorithms for Clustering Data (1988). He served as a member of the Defense Science Board and The National Academies committees on Whither Biometrics and Improvised Explosive Devices. He received the 1996 IEEE TRANSACTIONS ON NEURAL NETWORKS Outstanding Paper Award and the Pattern Recognition Society Best Paper Awards in 1987, 1991, and 2005. He has received Fulbright, Guggenheim, Humboldt, IEEE Computer Society Technical Achievement, IEEE Wallace McDowell, ICDM Research Contributions, and IAPR King-Sun Fu awards. He is a Fellow of the AAAS, ACM, IAPR, SPIE, and IEEE.

APPENDIX A
DATABASES FOR DEMOGRAPHIC ESTIMATION

A number of face databases have been used to study the demographic estimation problem [1]–[12]. However, many of these databases, such as the YGA [3], Asian dataset [4], and BERC [6], are not available in the public domain.

For facial aging and age estimation studies, the two most popular databases are FG-NET [1] and MORPH II [2]. For gender classification, the FERET database [9] is the most popular. However, there is no commonly used database for race classification. Our experiments on demographic estimation are based on these three public-domain databases, namely FG-NET, FERET, and MORPH II. We also use a large database, namely PCSO database, obtained from the Pinellas County Sheriff's Office. In addition, we used a subset of the LFW database [17] to evaluate the generalization ability of the proposed approach to the general population of faces.

The FG-NET database, one of the first publicly available face databases with real ages provided for each subject, has played an important role in advancing research on age estimation. However, as shown in Fig. 1 (a), the age distribution of this database is strongly biased to younger ages (<18 years). Additionally, the number of subjects in FG-NET is small (only 82 subjects), so it cannot be used effectively to design reliable age estimation algorithms. However, to compare the performance of our algorithm against published results, we follow a leave-one-person-out protocol on the FG-NET database.

The FERET database has been used for studying gender classification [13]–[16]. However, since the race distribution in FERET is biased significantly to whites (see Table 1), and the age distribution of subjects is highly concentrated to a few discrete ages (e.g., 20, 30, and 40), we only use FERET to study gender classification. We use the Color FERET database, which contains $2,719$ frontal face images (fa and fb), with a 5-fold Leave-One-Fold-Out (LOFO) protocol.

MORPH is a large database of mugshot images, each with associated metadata containing age, gender, and race information. We investigate all the three demographic estimation tasks on MORPH Album2 (MORPH II), with a commercial version of the database containing $78,207$ images of $20,576$ subjects that was released on February 2010. Results of MORPH II are reported with a 5-fold LOFO protocol.

The PCSO database, contains mugshot images with metadata, including the image capture date, date of birth, gender, and race[1]. The complete PCSO database available to us contains ~ 1.5 million mugshots, out of which we sample a subset of $100,012$ images of

1. Interested researchers may contact the PCSO to access this database.

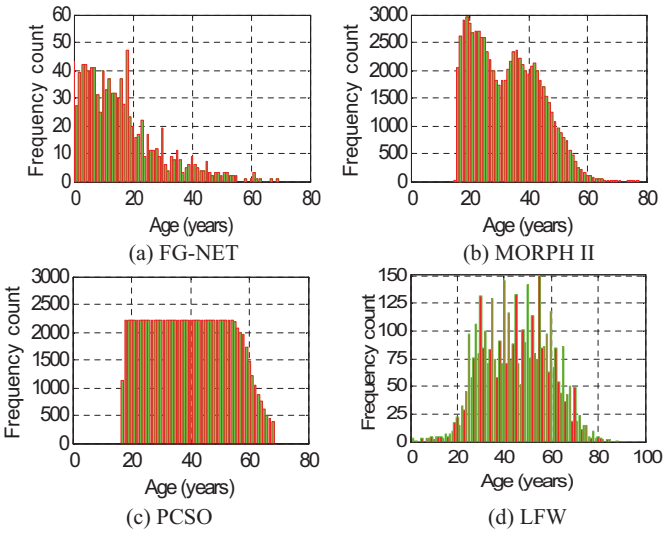


Fig. 1. Age distributions in (a) FG-NET (1,002 images), (b) MORPH II (78,207 images), (c) PCSO (100,012 images), and (d) a subset of LFW (4,211 images) databases.

$81,533$ subjects with near-uniform age distribution. Age estimation on PCSO demonstrates our capability in handling a near-uniform age distribution (see Fig. 1 (c)), which is more challenging, since we cannot sacrifice the accuracy in a particular age range to improve the overall performance, as in the case of FG-NET. Gender and race classifications are also performed on PCSO. Note that PCSO has a completely different race distribution than MORPH II, e.g., more whites than blacks. Results of PCSO are reported with a 5-fold LOFO protocol.

Unlike the previously discussed databases, the LFW database contains face images captured under unconstrained conditions. We perform age, gender, and race estimation on a subset of LFW with $4,211$ subjects (one image per subject), where the face images have relatively small pose variations (see Fig. 2 (e)). Since the real age, gender, and race for LFW images are not available, we collected human (crowdsourced) estimates of age, gender and race of each face image using Amazon Mechanical Turk (MTurk) with three workers per task. The appearance estimates of age, gender, and race are used as the ground truth to evaluate the proposed approach. Results of LFW are reported with a 5-fold LOFO protocol.

The five databases used in our experiments cover a wide variety of acquisition scenarios. FG-NET is comprised of personal photographs. FERET is collected from cooperative subjects. MORPH II and PCSO are two operational mugshot databases from law enforcement agencies. LFW represents general population of faces under unconstrained conditions. We summarize the characteristics of these five databases in Fig. 1 and Table 1. Example face images from these five databases are shown in Fig. 2.

TABLE 1
Gender and race distributions for MORPH II, PCSO, FERET, and LFW databases.

	# Images	MORPH II		PCSO		FERET	LFW
	# Subjects	Machine	Human	Machine	Human	Machine	Machine
Gender	Female	12,606 3,553	1,000 845	25,006 20,985	1,100 227	1,007 403	1,101 1,101
	Male	65,601 17,023	1,000 943	75,006 60,548	1,100 138	1,712 591	3,110 3,110
Race	White	15,996 4,660	1,000 856	69,116 56,660	1,000 615	1,689 618	3,501 3,501
	Black	58,326 14,405	1,000 932	26,457 20,742	1,000 451	215 78	352 352
	Other	3,885 1,511	0 0	4,439 4,131	0 0	815 298	358 358
Total		78,207 20,576	2,000 1,788	100,012 81,533	4,200 1,431	2,719 994	4,211 4,211

Machine and Human denote the datasets used in demographic estimations by the proposed approach and human (MTurk workers), respectively. Statistics of the LFW databases are based on the human estimates by MTurk workers.

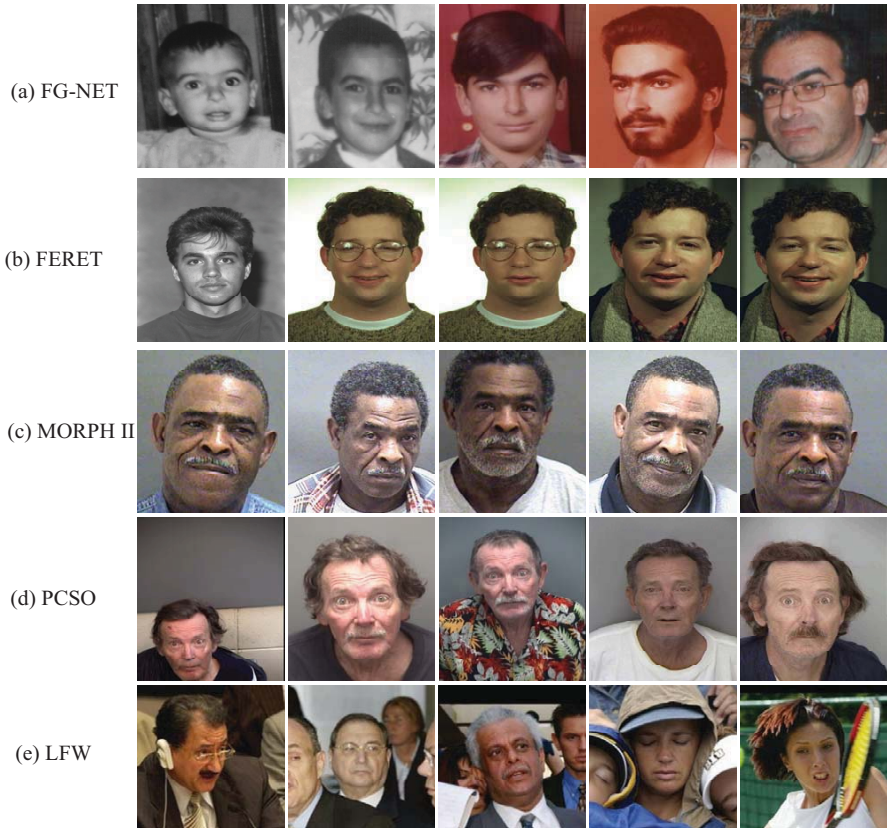


Fig. 2. Example face images from the (a) FG-NET, (b) FERET, (c) MORPH II, (d) PCSO, and (e) LFW databases.

APPENDIX B
FACE IMAGE QUALITY ASSESSMENT

Quality assessment is learned in the training stage, and then utilized in the testing stage. We partition the training set into high-quality (positive) and low-quality (negative) subsets. Taking age estimation as an example, if 3% (P) face images in the training set with the largest absolute error of age estimation are used as low-quality (negative) samples, the remaining 97% ($1-P$) face images in the training set will be the high-quality (positive) samples (see Fig. 3 (a)). Figure 4 shows examples of low-quality face images detected by the proposed quality assessment model.

APPENDIX C
DEMOGRAPHIC ESTIMATION BY HUMAN

In our experiments on demographic estimation by human, the GUI constrains the user's possible input for age, gender, and race estimation tasks. Specifically, we do not allow MTurk workers to input an age range; otherwise, the task would be *age group* rather than *exact age* estimation, which is studied in this work. Only a binary choice is allowed for gender (female vs. male) and race (black vs. white)[2] classifications. The main purpose of above constraints is to provide

2. We allow more options for race classification in the LFW database, *i.e.*, Black, White, Asian, and Unknown, but these human estimates of age, gender and race are used as the ground truth, not for the purpose of comparisons between human and machine performance.

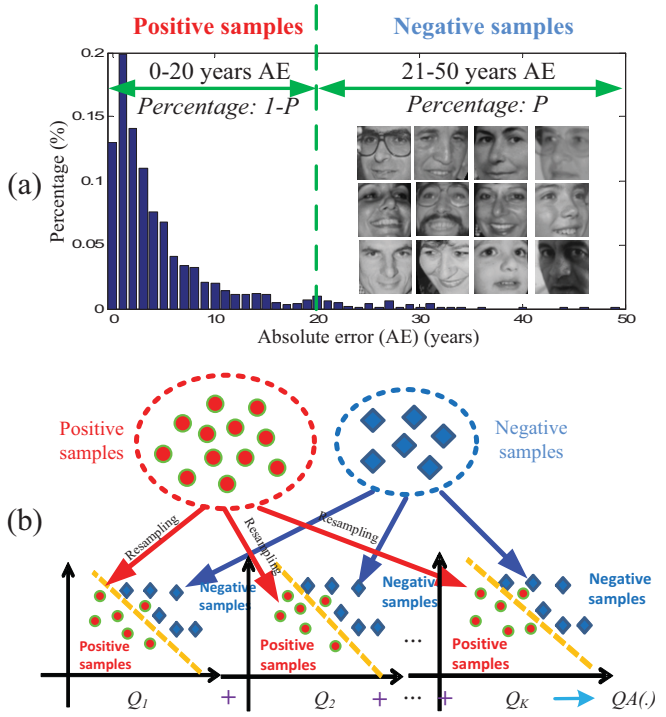


Fig. 3. Learning-based quality assessment for demographic estimation: (a) positive (high-quality) and negative (low-quality) samples from the FG-NET database, and (b) an ensemble quality assessment classifier consisting of multiple binary SVM classifiers learned with resampling of positive samples.

Fig. 4. Examples of low-quality (large pose and expression variations, overexposure, and out of image facial regions) face images detected by the proposed quality assessment method in the MORPH II database.

a fair comparison between the proposed demographic estimation method and human. As shown in Table 1, the number of subjects of "the other race" is much smaller than those of black and white subjects, which restricts the training of the classifier for "the other race". However, humans have been well "trained" to distinguish different race groups in their daily life. If we consider three race groups (black, white, and other race), a fair comparison between human and our algorithm may not be possible. The footnote under Table 3 in the main manuscript explains the race groups (such as Asian, Caucasian, and African) considered by some of the published methods, but no human performance is reported on these race groups. Similarly, for gender classification, since the face images submitted to MTurk only contain male and female subjects, we believe it is reasonable to ask the MTurk workers to make a binary decision: male or female.

We also notice that the absolute age estimation errors by human and machine are typically within 30 years on the FG-NET, MORPH II, and PCSO databases. One may question the accuracy of the real ages provided in the databases if significant age estimation errors (by either human or machine) are observed (e.g., >30 years). Figure 5 shows two examples from the PCSO database where human estimation errors are larger than 30 years, but obviously the real (labeled) ages provided in the database are incorrect.

In the comparisons between human and machine performance, it is reasonable to compare human performance with machine performance with *cross-database* testing, because human performance is not tuned to a particular database. We have also applied quality assessment on the datasets used for reporting human performance. Human performance is reported with and without quality assessment, which is exactly the same as how machine performance is reported. There are some studies in literature where human vs. machine performance is reported on face recognition. Most of these publications were reviewed in [18], [19]. In this paper, we focus on the study of human vs. machine performance on demographic estimation.

Another interesting problem in studying the human perception ability to demographics would be the other-race effect, such as the East Asian's performance on Caucasian faces, and vice versa. There is still some limitations in doing such studies via MTurk because a recent study of human performance on face recognition [19] by MTurk workers shows that most of the MTurk workers (totally 307 unique workers who provided $\sim 60,000$ responses) are from India (55.1%) and US (27.4%), so the demographics of MTurk workers may not be ideal for studying specific cases such as East Asian's performance on Caucasian faces. In our future work, we may restrict the country of origin of the MTurk workers to obtain the required human responses from a particular race group or a country of origin.

References

[1] Facial Image Processing and Analysis (FIPA), "FG-NET Aging Database," http://fipa.cs.kit.edu/433.php#Downloads.

[2] K. Ricanek and T. Tesafaye, "MORPH: A Longitudinal Image Database of Normal Adult Age-progression," *Proc. FGR*, pp. 341-345, 2006.

Age in the database: 16
Proposed: 40, Human: 55

Age in the database: 17
Proposed: 58, Human: 53

Fig. 5. Examples of incorrect real ages of two subjects provided in the PCSO database.

[3] Y. Fu and T.S. Huang, "Human Age Estimation with Regression on Discriminative Aging Manifold," *IEEE Trans. Multimedia*, vol. 10, no. 4, pp. 578-584, Jun. 2008.

[4] J. Suo, T. Wu, S.-C. Zhu, S. Shan, X. Chen, and W. Gao, "Design Sparse Features for Age Estimation Using Hierarchical Face Model," *Proc. FGR*, pp. 1-6, 2008.

[5] M. Minear and D. Park, "A Lifespan Database of Adult Facial Stimuli," *Behav. Res. Meth. Ins. C.*, vol. 36, no. 4, p. 630-633, Nov. 2004.

[6] S.E. Choi, Y.J. Lee, S.J. Lee, K.R. Park, and J. Kim, "Age Estimation Using a Hierarchical Classifier Based on Global and Local Facial Features," *Pattern Recogn.*, vol. 44, no. 6, pp. 1262-1281, Jun. 2011.

[7] T. Wu, P. Turaga, and R. Chellappa, "Age Estimation and Face Verification across Aging Using Landmarks," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 6, pp. 1780-1788, Dec. 2012.

[8] A. Hadid and M. Pietikänen, "Demographic Classification from Face Videos Using Manifold Learning," *Neurocomputing*, vol. 100, pp. 197-205, Jan. 2013.

[9] P.J. Phillips, H. Moon, S. A. Rizvi, P.J. Rauss, "The FERET Evaluation Methodology for Face-recognition Algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 10, pp. 1090-1104, Oct. 2000.

[10] X. Lu, H. Chen, and A.K. Jain, "Multimodal Facial Gender and Ethnicity Identification", *Proc. ICB*, pp. 554-561, 2006.

[11] C. Chen and A. Ross, "Evaluation of Gender Classification Methods on Thermal and Near-infrared Face Images," *Proc. IJCB*, pp. 1-8, 2011.

[12] D. Cao, C. Chen, M. Piccirilli, D. Adjeroh, T. Bourlai, and A. Ross, "Can Facial Metrology Predict Gender?" *Proc. IJCB*, pp. 1-8, 2011.

[13] B. Moghaddam and M. Yang, "Gender Classification with Support Vector Machines," *Proc. FGR*, pp. 306-311, 2000.

[14] S. Gutta, J.J. Huang, P. Jonathon, and H. Wechsler, "Mixture of Experts for Classification of Gender, Ethnic Origin, and Pose of Human Faces," *IEEE Trans. Neural Netw.*, vol. 11, no. 4, pp. 948-960, Jul. 2000.

[15] Z. Yang and H. Ai, "Demographic Classification with Local Binary Patterns," *Proc. ICB*, pp. 464-473, 2007.

[16] J.E. Tapia and C.A. Perez, "Gender Classification Based on Fusion of Different Spatial Scale Features Selected by Mutual Information from Histogram of LBP, Intensity, and Shape," *IEEE Trans. Inf. Forensics Security*, vol. 8, no. 3, pp. 488-499, Mar. 2013.

[17] G.B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments," *University of Massachusetts, Amherst, Technical Report 07-49*, Oct. 2007.

[18] N. Kumar, A. Berg, P. Belhumeur, and S. Nayar, "Describable Visual Attributes for Face Verification and Image Search," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 10, pp. 1962-1977, Oct. 2011.

[19] L. Best-Rowden, S. Bisht, J. Klontz and A. K. Jain, "Unconstrained Face Recognition: Establishing Baseline Human Performance via Crowdsourcing," *Proc. IJCB*, Sept. 2014.