**Noname manuscript No.**
(will be inserted by the editor)

# Real-Time Demographic Profiling from Face Imagery with Fisher Vectors

**Lorenzo Seidenari · Alessandro Rozza · Alberto Del Bimbo**

Received: ... / Accepted: ...

**Abstract** In the last decade, demographic profiling from facial imagery has grown in its importance in the computer vision field. For demographic profiling, we usually mean gender, ethnicity, and age identification from face images. In this paper, we propose an efficient and effective profiling framework and we assess the quality of the proposed approach comparing the results obtained by our system with those achieved by other recently published methods on large datasets of facial images with different age, gender, and ethnicity. These results show how a carefully engineered pipeline of efficient image analysis and pattern recognition techniques leads to state-of-the-art results at 20FPS using a single thread on a 1.6GHZ i5-2467M processor.

**Keywords** Demographic Face Profiling · Age Estimation · Gender Classification · Ethnicity Classification

## 1 Introduction

The analysis of face imagery offers the possibility to identify many properties at different levels of specificity. Some of the most interesting are: gender, ethnicity, and age. In this work, we will consider the union of these three properties as demographic profiling. More formally, the demographic face profiling task can be defined as, given one or more subsequent samples of face images, to obtain a single prediction on a set of attributes (age, gender and ethnicity).

In the last decade, demographic profiling from facial imagery has grown its importance in the computer vision field. The process of gender, ethnicity, and age determination (disjoint, partially joint, or joint) finds several application areas. A person's age could be verified to implement age-based access control and verification, prior to physical access to a place or product being sold or virtual access to a website is granted.

In the task of targeted advertising, as an example, a digital sign can display commercials based on demographics of audience walking past. Soft biometrics makes use of ethnicity, gender, and age-based to index face images into huge-scale biometric databases for faster retrieval.

Furthermore, the analysis of crowded environments to identify the age, the gender, and the ethnicity distributions of the people is becoming strategic for the retail chains. All the aforementioned applications have real-time requirements, since attribute estimation must be performed as fast as possible, and in certain cases on devices with limited resources. The severity of this requirement increases, especially if there is a limited amount of time to make a decision. Taking as an example the targeted advertising case, the temporal window coincides with the time a person takes to walk past the digital sign.

Face imagery exhibits many variations which may affect the ability of a computer vision system to infer such attributes. We can categorize these variations as being caused by the human or the image capture process. Factors related to the image capture process are the head pose, the illumination, and the image quality (blurring, noise, low resolution). Human factors are due to the characteristics of a person, such as facial expressions (surprise, neutral, happy/smiling, etc.), and the accessories being worn (eye glasses, hat, etc.).

In real-world scenarios, a fixed camera will acquire several shots of a person's face from a video stream.

Lorenzo Seidenari and Alberto Del Bimbo
University of Florence
E-mail: {lorenzo.seidenari, alberto.delbimbo}@unifi.it

Alessandro Rozza
E-mail: alessandro.rozza.it@ieee.org

This setting introduces an interesting potential for improving attribute estimation accuracy, but also another set of problems. Indeed, the final estimate can be computed over a sequence of predictions. Although the availability of multiple face shots for each target introduces a profile consistency requirement.

In case single person profiling is sought, the algorithm must provide a one-to-one profile-person correspondence, in order to output useful information. Also, in the case where demographics of people are collected, this correspondence is very relevant. Indeed, attribute statistics may be influenced by the time a person persist into the camera field of view, if predictions are not grouped in a single profile.

First of all, detected faces must be associated across different frames, maintaining identity information so to enforce profile consistency. Moreover, to also obtain an advantage, an image quality metric must be devised to weight single-frame predictions. Indeed many errors are often due to extreme face poses and occlusions.

### 1.1 Contribution

We propose a thoroughly and carefully engineered computer vision and image processing pipeline for demographic profiling, which is suitable for real-time embedded environments. We extend our previous contribution [53], in many ways. First, we approach the full demographic profiling problem, by predicting age, ethnicity and gender. Ethnicity and gender predictors are often learned on imbalanced datasets. We therefore apply Truncated Isotropic Principal Component Analysis Classifier (`TIPCAC` [50]) which allows us not to perform re-sampling on data. Second, we present a novel method to incorporate semantic predictions from video sequences. Finally, we report results on all three tasks on two large scale datasets.

## 2 Related Work

Gender estimation is usually performed as a binary classification problem [43]. Linear and non-linear classifiers are employed using different kind of features such as Local Binary Pattern (LBP), Gabor Wavelet, and Scale Invariant Feature Transform (SIFT). LBP are often used as local features to predict Gender. In [5], the authors exploit LBP with Support Vector Machines (SVM) for multi-view gender classification. Eidinger *et al.* [15] propose to extract four LBP patches and to weight the reliability of each patch using a probabilistic model to assess landmark estimation accuracy. A more advanced method for gender estimation has been proposed by Hassner *et al.* [29], which employes LBP and a novel frontalization approach. In [59], the authors extract SIFT descriptors combined with global shape contexts of the face and they perform the final classification using Adaboost. Gabor filters have been used to obtain the simple cell units of the Biologically Inspired Features (BIF) proposed by Riesenhuber and Poggio [48] for object recognition and later extended by Meyers and Wolf [41] for face processing and by Guo *et al.* [20] for gender estimation. In [10], Cirne and Pedrini propose a method for gender recognition based on a novel geometric descriptor built on a pre-defined face shape model. Precisely, this technique extracts 68 fiducial points of the face shape and it computes the pairwise Euclidean distances to obtain a descriptor of 2278 values. This approach, compared with other well-known geometric methodologies, achieves promising results.

Compared to gender identification, ethnicity classification has received less attention by the computer vision community. Ethnicity estimation is usually performed as a multi-class classification problem. Since an ethnic group or ethnicity is a category of people who identify with each other based on common ancestral, social, cultural or national experience[1] many possible ethnicity classes can be identified. To simplify the problem, it is recommended that the maximum amount is less than 10. One of the most important problem in ethnicity detection is the unbalanced characterization of the standard datasets. This problem can affect the quality of the final performance, since many classification techniques are not able to deal with this specific task. Another important problem is related to the fact that many ethnic groups have strong visual similarity between them, thus increasing the overlap between the features that describe the classes and enhancing the complexity of the classification.

Guo *et al.* conducted a large scale study on the MORPH-II dataset analyzing the influence of gender and age on the prediction of ethnicity [21]. They performed an experiment using a balanced subset with only White and Black subjects and also a more comprehensive one with the whole dataset. They showed that to develop an effective ethnicity estimation method is a challenging task given the scarcity of annotated data and the extreme bias of the existing ones.

Age estimation is the most investigated of the considered three tasks. It is usually performed as a multi-class classification or as a regression problem. In the first case, the age labels are quantized in a set of age groups, e.g. $\{[16, 25], [26, 35] \ldots [56, 65]\}$. Instead, in the

[1] "ethnicity: definition of ethnicity". Oxford Dictionaries. Oxford University Press. Retrieved 28 December 2013.

regression problem, the age is treated as a real number and a function is computed to minimize the age estimation error. This approach has several advantages over the multi-class classification task. First, the overall data can be used to fit a single model. Second, the loss function can be formulated more naturally penalizing models proportionally to the error they commit.

Some interesting related works in age estimation exploit BIF. BIF are firstly proposed for age estimation by Guo *et al.* [24] combined with a linear SVM. In their work the authors use a pyramid of Gabor filters with small sizes and they suggest to determine the number of orientations and bands with an ad-hoc approach. In [27] the author present a method that identifies different facial components and extracts BIF features describing these parts. Each component is classified into one of four disjoint age groups using a decision tree and a final regressor is trained to compute the age.

Chang *et al.* in [9] proposed an ordinal hyperplane ranker on Active Appearance Models (AAM, [11]) that exploits the distribution of the training labels. The key idea is try to obtain multiple decisions on who is the older of two people to finally determine the person's actual age.

In [30], Hou et al. present a novel loss function that is able to better capture inter-class relationships clearly present in age estimation as also in other real-life tasks. In this work, they propose to train deep neural networks with the exact squared Earth Movers Distance (also called Wasserstein distance). This loss uses the predicted probabilities of all classes and penalizes the miss-predictions according to a ground distance matrix that quantifies the dissimilarities between classes.

Geng *et al.* propose two methods that exploit the label distributions [19] of the face imagery. Instead of considering each facial image as an instance with a single age value, the authors consider each image as associated with a label distribution. The label distribution covers a certain number of class labels. In this way, this approach guarantees that one face image can contribute also to the learning phase of its adjacent ages.

Fu and Huang [18] proposed an age estimation framework that is composed by two main modules, a discriminative manifold learning approach followed by a multiple linear regression. Precisely, the proposed manifold learning algorithm (Conformal Embedding Analysis, CEA) is a supervised subspace learning method, which incorporates the labeling information of both neighborhood and class to each sub-manifold. Moreover, the age estimation is performed as a multiple linear regression problem in the manifold space. The authors have experimentally shown that the better results are achieved using a quadratic model function.

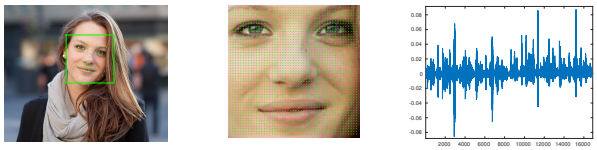
Ni et al. [46] present an automatic image and video mining framework with the aim of building a cross-ethnicity human age estimator based on facial information. To achieve this goal the authors propose a robust multi-instance regressor learning algorithm able to deal with images with multiple face instances and possibly noisy images and labels. To train their approach they collect a large size human aging image dataset from Flickr and Google Image.

Guo *et al.* propose to employ the kernel partial least squares regression (KPLS) for age estimation [22]. The strength of this approach is that KPLS simultaneously performs the feature dimensionality reduction and learns the aging function improving the final results in terms of accuracy and reducing the time cost.

A comprehensive list of recent age estimation approaches can be found in [27].

There is very little work on face attribute estimation from video. DeMirkus *et al.* [12] proposed a hierarchical Bayesian method, but only addresses gender estimation on video sequences. Their approach exploits a multi-part face representation and a temporal modeling to deal with the challenges generated by partial occlusions, expression and pose variations, present in unconstrained video sequences.

There are very few papers focusing on complete demographic profiling. In some works ethnicity and gender estimation are treated as a pre-processing to improve age estimation, learning more specialized predictors. As already highlighted in this section, in [20], [21] Guo *et al.* investigate the variations of age estimation performance under variations across race and gender. They observe that crossing race and gender can result in significant error increases for age estimation. To leverage the aging pattern of different gender and ethnicity, they employ the feature presented in their previous work [24] and they propose a 3-step method learning separate classifiers for different combinations of age and genders and applying the age estimator only after predicting the gender and ethnicity of the subject. In [36], Lapuschkin et al. compare four popular neural network architectures (CaffeNet, GoogleNet, VGG-16, and AdienceNet [37]), study the effect of pretraining, and evaluate the robustness of different alignment preprocessing, on gender and age estimation. Moreover, by employing Layer-wise Relevance Propagation the authors investigate which facial features are actually used for age and gender prediction. In [61], Zhang et al. introduce a very deep neural network architecture for age group and gender estimation leveraging Residual Networks of Residual Networks (RoR, [62]). To reduce the overfitting problems and to increase the performance, the RoR model is pre-trained on ImageNet, fune-tuned



(a) Landmark estimation (b) Feature Sampling (c) Fisher Vector

Fig. 1: Our image representation pipeline. Face detection and landmark estimation (a) followed by dense multi-scale SIFT extraction on the aligned face (b) and Fisher Vector computation (c).

on the IMDB-WIKI-101 dataset and (finally) on Adience dataset.

More recently, a deep convolutional neural network has been designed to handle the age and gender profiling problem by Levi *et al.* [37]. In [13], Duan et al. extends their previous work [14] to propose a methodology to ensemble Convolutional Neural Networks (CNNs) and extreme learning machine (ELM, [31]) to perform age estimation. Precisely, this approach combines, in a hierarchical fashion, three CNNs (called Age-Net, Race-Net, and Gender-Net) with ELM classifier and ELM regressor. The CNNs are used to extract features, while the final ELM regressor predicts age.

To the best of our knowledge, our work is the only one addressing the whole demographic profiling spectrum, also dealing with video sequences. Moreover, we attain real-time performance without specialized hardware such as GPUs or ASICs. In this paper, we describe our demographic profiling system designed with efficiency in mind. Different from previous works we use a high-dimensional modern feature [57] (see Fig. 1) that proves to be accurate yet efficient for age, gender and ethnicity identification.

This paper is organized as follows: in Section 3 our face representation is summarized; in Section 4 the employed face detection approach, the tracking method to maintain associated the detected face, and the alignment technique are described; in 5 the approaches used for gender, ethnicity, and age estimation are presented; in Section 6 the achieved results on large datasets are shown; in Section 7 our conclusions are highlighted.

## 3 Face Representation

We design our face representation inspired by recent results in image classification [54] and face recognition [56]. We build on our previous contribution on age estimation [53]. We assume our face patches are aligned to a common, fixed size, reference square. This assumption is easily satisfied by our face extraction pipeline described in Sec. 4.

As a local feature we use densely sampled SIFT descriptors [39]. We compute descriptors at multiple fixed scales, discarding orientation estimation. The use of multiple scales helps in representing facial features that may appear at different sizes, even on fixed size face patches. Since we are computing the representation for aligned faces, we can exploit not just the local image statistics but also local feature coordinates. SIFT descriptor are pre-processed using PCA and preserving 64 components and then augmented with their x,y coordinates, rescaled in $[-1, 1]$.

To obtain a global face descriptor, we apply Fisher Vector encoding to the aforementioned compressed and augmented SIFT features. Fisher Vectors require a Gaussian Mixture Model dictionary to be computed. We learn this dictionary, and the PCA transformation, on a set of 200K randomly drawn SIFTs from the training set.

Given a Gaussian Mixture Model dictionary with parameters $\boldsymbol{\mu}_n, \boldsymbol{\sigma}_n, \boldsymbol{\omega}_n$ and given soft-assignments $\gamma_m^{(n)}$ for each of the $M$ augmented SIFT feature $\boldsymbol{x}_m \in \boldsymbol{X}$, the Fisher Vector is computed concatenating the following gradients:

$$\mathcal{G}_n^{\mu}(\boldsymbol{X}) = \frac{1}{\sqrt{\boldsymbol{\omega}_n}} \sum_{m=1}^{M} \gamma_m^{(n)} \left( \frac{\boldsymbol{x}_m - \boldsymbol{\mu}_n}{\boldsymbol{\sigma}_n^2} \right), \tag{1}$$

$$\mathcal{G}_n^{\sigma}(\boldsymbol{X}) = \frac{1}{\sqrt{2\boldsymbol{\omega}_n}} \sum_{m=1}^{M} \gamma_m^{(n)} \left( \frac{(\boldsymbol{x}_m - \mu_n)^2}{\boldsymbol{\sigma}_n^2} - 1 \right), \tag{2}$$

where

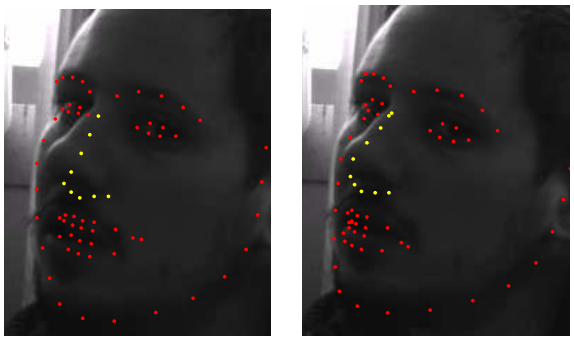
$$\gamma_m^{(n)} = \frac{\boldsymbol{\omega}_n p_n(\boldsymbol{x}_m)}{\sum_{j=1}^{D} \boldsymbol{\omega}_j p_j(\boldsymbol{x}_m)}, \tag{3}$$

and $p_n$ is the $n^{th}$ Gaussian of the learned mixture and $\mathbf{X}$ is the feature set of a face image. The size of our descriptor is $F \times D \times 2$, where $F$ and $D$ are local feature and dictionary size. As an example, if we consider $D = 128$, since $F = 66$ our face signature size is 16896.

## 4 Face Detection, Tracking and Alignment

Our method performs profiling over a set of aligned face patches. Alignment is required to compute the face representation described in Sec. 3, where feature coordinates, in a common reference, are exploited to improve appearance descriptors. To gather a set of face patches, we design an efficient processing pipeline.

In this section, we describe our image processing pipeline, that precedes the face profiling step. As a first



(a) Landmarks estimated without equalization.

(b) Landmarks estimated with equalization.

Fig. 2: Face landmark detection without (a) and with (b) equalization on a challenging image. Nose landmarks, marked in yellow, are wrongly localized without equalization.

step, we apply a low-level image pre-processing to improve the successive face detection and alignment steps. For every detected face we compute a transformation obtaining a consistent geometric reference for image features. Moreover, a multi-target tracker is employed to maintain association between different frames when dealing with video sequences, instead of single facial imagery.

### 4.1 Image Pre-Processing

Face misses are a critical issue, since they reduce the effectiveness of our method. Face detection may fail in case of highly saturated images, that may happen, for example, when a camera is facing the entrance of a building.

While often face normalization has the goal to locally normalize a face in order to obtain invariance to illumination changes, we are mainly interested in compensating sensor saturation in presence of strong lighting. Our goal is to detect as many faces as possible without compromising real-time performance. Considering our task and the real-time constraint we evaluated rank and wavelet based normalization [55]. In a set of preliminary experiments we found out that a basic histogram equalization, is enough to improve face detection recall and landmark localization.

In Fig. 2, it can be seen that detecting landmarks in a highly saturate image, without any pre-processing, results in poor accuracy. It can be noted that in Fig. 2a, nose landmarks are wrongly localized, while in Fig. 2b they have a more consistent location.

### 4.2 Face detection

Face detection is the most expensive step of our pipeline, it requires an exhaustive, multi-scale, search over examined frames. To avoid trading accuracy for efficiency, we apply a very simple yet effective linear classifier. The model is trained with structural SVM on $\sim$3000 faces. Five face poses are considered in training: frontal, profile-left, profile-right, frontal left-tilted and frontal right-tilted. The structural SVM formulation of [35] is efficient to train and obtains state-of-the art results even with linear classifiers.

### 4.3 Face alignment

We apply a face alignment step to gain invariance to face pose. This step is extremely important, since our feature is computed from the joint statistic of intensity and location of pixels. Faces are aligned and rescaled to a common reference frame. Affinity based alignment is used, which performs a non-uniform scaling along the two dimensions. We estimate the rotation, translation and scaling matrix mapping the eye-mouth triangle to a canonical triangle: $(0.2 \cdot S, 0.2 \cdot S)$, $(0.8 \cdot S, 0.2 \cdot S)$, $(0.5 \cdot S, 0.5 \cdot S)$ where $S$ is the square size. As highlighted by Fig. 3(b) all important facial features can be recovered, which is not always possible using a simpler rigid rotation.

A cascade of regression trees is used to estimate the face shape. Trees are trained on pixel intensities and, for each detected face, extract 68 landmarks[34]. Eyes and mouth centers are robustly estimated using the median of 20 and 6 landmarks for eyes and mouth respectively. Faces are then remapped in a $100 \times 100$ pixel rectangle using the aforementioned affine tranform.

Using our alignment pipeline. we can effectively deal with ($\pm 15°$) yaw variations, in case of higher pose variations, full 3D approach ought to be used [29]. Unfortunately 3D frontalization is computationally expensive.

### 4.4 Face Tracking

We use a greedy association multi-target tracker. At each frame a set of face detections $\mathcal{D}_t$ is generated applying the multi-pose face detector described in Sec. 4.2. Considering the, possibly empty, set of tracks $\mathcal{T}_{t-1}$ present at the previous frame, we compute an association matrix $\mathbf{M}$ such that $\mathbf{M}_{ij} = \frac{d_i \cap t_j}{d_i \cup t_j}$ also known as the intersection over union measure. To track a person face, we apply the function associate($\cdot$) described in Algorithm 1.

(a) Rotation alignment. (b) Affine transform alignment.

Fig. 3: Alignment results with rotation compensation and with affine alignment. In the face marked in red the mouth is missing in the rotated image whilst using the affine compensation all important facial features are visible.

**FUNCTION** associate $(\mathcal{T}_{t-1}, \mathcal{D}_t)$
**Data**: $\mathcal{T}_{t-1} : \{t_1 \dots t_n\}, \mathcal{D}_t : \{d_1 \dots d_m\}, \mathbf{M}_{ij} = \frac{d_i \cap t_j}{d_i \cup t_j}$
**Result**: $\mathcal{T}_t$
**while** $\max_{ij} \mathbf{M}_{ij} > \tau$ **do**
  **if not** $\mathbf{A}_{ij} \wedge \mathbf{M}_{ij} > \tau$ **then**
    $\langle \hat{i}, \hat{j} \rangle \leftarrow \arg\max_{ij} \mathbf{M}_{ij}$;
    $t_{\hat{i}} \leftarrow d_{\hat{j}}$ $\mathbf{A}_{\hat{i}:} \leftarrow$ `TRUE`;
    $\mathbf{A}_{:\hat{j}} \leftarrow$ `TRUE`;
  **end**
**end**
`/* Unassigned detections initialize new tracks. */`
$\mathcal{T}_t \leftarrow \mathcal{T}_{t-1} \cup \{d | \mathbf{A}_{ij} = \texttt{TRUE}\}$;
`/* Remove tracks not assigned for` $\omega$ `frames. */`
$\mathcal{T}_t \leftarrow \mathcal{T}_{t-1} \setminus \{t_i | \mathbf{l}_i > \omega\}$

**Algorithm 1:** Data association algorithm. We associate tracks and unassociated detection if IoU $> \tau$ and remove a track if it is "dead" for $\omega$ frames. Matrix $\mathbf{A}$ keeps track of associations and vector $\mathbf{l}$ counts the amount of frames a track $i$ is not associated with any detection.

## 5 Profiling

In this section we present the approaches employed to identify the gender, the ethnicity, and the age of the analyzed subjects. Since these problems are respectively a binary classification problem, a multi-class classification problem, and a regression/multi-class classification task, we have exploited different methodologies to improve the quality of the final results. Precisely, in Section 5.1 the approach for the gender classification is described; in Section 5.2 the approach to estimate the ethnicity is proposed; in 5.3 the methods for the age estimation are formalized. Finally, in 5.4 we describe how to apply our algorithm in real world scenarios exploiting video sequences.

### 5.1 Gender Estimation

In this section we describe the employed binary classification method for gender identification. This approach is particularly suitable when it is employed on binary classification problems with high dimensional data and when the distribution, that underlines the points, can be well approximated by a Mixture of Gaussians (as in our case).

The base implementation of `T-IPCAC`, called Isotropic Principal Component Analysis Classifier (`IPCAC`), has been presented in [51]. Given a set of $N$ clustered points sampled from an isotropic Mixture of Gaussians, the Fisher subspace (`Fs`) corresponds to the span of the class means; as a consequence, when a binary classification problem is considered, `Fs` is spanned by unit vector $\boldsymbol{f} = \frac{\boldsymbol{\mu}_A - \boldsymbol{\mu}_B}{||\boldsymbol{\mu}_A - \boldsymbol{\mu}_B||}$, where $A$ and $B$ are the two classes, and $\boldsymbol{\mu}_{A/B}$ the class means.

`IPCAC` exploits this result by whitening[2] the training set $\boldsymbol{\mathcal{P}}_{Train} = \{\boldsymbol{p}_1 \dots \boldsymbol{p}_N\}$, computing $\boldsymbol{f}$, and classifying

[2] We call "white" a dataset of points sampled from a probability distribution with $\boldsymbol{\mu} = \mathbf{0}$, and $\boldsymbol{\Sigma} = \boldsymbol{I}$ where $\boldsymbol{I}$ is the identity matrix.

a new point $\boldsymbol{p}$ as follows:

$$\theta((\boldsymbol{W}_D^T \boldsymbol{f}) \cdot \boldsymbol{p} - \gamma) = \theta(\boldsymbol{w} \cdot \boldsymbol{p} - \gamma);$$
$$\gamma = \bar{\mu}_A + \frac{\sigma_A(\bar{\mu}_B - \bar{\mu}_A)}{\sigma_A + \sigma_B} \quad (4)$$

where $\theta(x) = A$ if $x \geq 0$, $\theta(x) = B$ if $x < 0$, the matrix $\boldsymbol{W}_D$ represents the whitening transformation estimated on the $N$ training points, $\bar{\mu}_A = \boldsymbol{w} \cdot (\boldsymbol{\mu}_A - \boldsymbol{\mu})$, $\bar{\mu}_B = \boldsymbol{w} \cdot (\boldsymbol{\mu}_B - \boldsymbol{\mu})$, $\boldsymbol{\mu}$ is the sample mean, and $\sigma_A$ and $\sigma_B$ are the standard deviations of the whitened points projected on $\boldsymbol{w}$.

Unfortunately our classification task is characterized by a training-set cardinality almost equal to the (high) space dimensionality. Under this setting the aforementioned technique often fails obtaining low quality results. To overcome this limitation, `T-IPCAC` [50] [52] improves `IPCAC` by replacing the first step of data whitening by a 'partial whitening' process. Precisely, if the points to be classified belong to a $D$ dimensional space, this method whitens the data in the linear subspace $\pi_d = \mathbf{Span}\langle \boldsymbol{v}_1, \cdots, \boldsymbol{v}_d \rangle$, spanned by the first $d \ll D$ principal components, while maintaining unaltered the information related to the orthogonal subspace $(\pi_d)^\perp = \mathbf{Span}\langle \boldsymbol{v}_{d+1}, \cdots, \boldsymbol{v}_D \rangle$.

Precisely, the linear transformation $\boldsymbol{W}_D$ is estimated as follows. The Truncated Singular Value Decomposition [28] is employed to estimate the first $d << D$ principal components[3], obtaining the low-rank factorization $\boldsymbol{P} \simeq \boldsymbol{U}_d \boldsymbol{Q}_d \boldsymbol{V}_d^T$ (where $\boldsymbol{P}$ is the matrix representing the training set $\boldsymbol{\mathcal{P}}_{Train}$ since it contains the training vectors). The $d$ largest singular values on the diagonal of $\boldsymbol{Q}_d$, and the associated left singular vectors, are employed to project on the subspace $\boldsymbol{\mathcal{SP}}_d$, spanned by the columns of $\boldsymbol{U}_d$, and to perform the whitening on the points contained in $\boldsymbol{P}$:

$$\bar{\boldsymbol{P}}_{\boldsymbol{W}_d} = q_d \boldsymbol{Q}_d^{-1} \boldsymbol{P}_{\perp SP_d} = q_d \boldsymbol{Q}_d^{-1} \boldsymbol{U}_d^T \boldsymbol{P} = \boldsymbol{W}_d \boldsymbol{P} \quad (5)$$

where $q_d$ is the smallest singular value of the points projected in $\boldsymbol{\mathcal{SP}}_d$. To obtain points whose covariance matrix best resembles a multiple of the identity, the value of the $d$ largest singular values is set to $q_d$ instead of 1, thus avoiding the gap between the $d$-th and the $(d+1)$-th singular value. The obtained matrix $\boldsymbol{W}_d$ projects and whitens the points in the linear subspace $\boldsymbol{\mathcal{SP}}_d$; however, dimensionality reduction during the whitening estimation might delete discriminative information, decreasing the classification performance. To avoid this information loss, this approach adds to the partially whitened data the residuals $\boldsymbol{R}$ of the points in $\boldsymbol{P}$ with respect to their projections on $\boldsymbol{\mathcal{SP}}_d$:

$$\boldsymbol{R} = \boldsymbol{P} - \boldsymbol{U}_d \boldsymbol{P}_{\perp \boldsymbol{\mathcal{SP}}_d} = \boldsymbol{P} - \boldsymbol{U}_d \boldsymbol{U}_d^T \boldsymbol{P} \quad (6)$$

$$\bar{\boldsymbol{P}}_{\boldsymbol{W}_D} = \left( q_d \boldsymbol{U}_d \boldsymbol{Q}_d^{-1} \boldsymbol{U}_d^T + \boldsymbol{I} - \boldsymbol{U}_d \boldsymbol{U}_d^T \right) \boldsymbol{P} = \boldsymbol{W}_D \boldsymbol{P} \quad (7)$$

[3] $d$ is a parameter to be set. Usually a good value is $d \simeq min(log_2^2 N, D)$

where $\boldsymbol{W}_D \in \Re^{D \times D}$ represents the linear transformation that whitens the data along the first $d$ principal components, while keeping unaltered the information along the remaining ones.

`Fs` is estimated by exploiting the whitened class means, $\boldsymbol{\mu}_A$ and $\boldsymbol{\mu}_B$, obtained by the class means estimated in the original space $\hat{\boldsymbol{\mu}}_A$ and $\hat{\boldsymbol{\mu}}_B$ as follows:

$$\boldsymbol{\mu}_A = q_d \boldsymbol{U}_d \boldsymbol{Q}_d^{-1} \boldsymbol{U}_d^T \hat{\boldsymbol{\mu}}_A + \hat{\boldsymbol{\mu}}_A - \boldsymbol{U}_d \boldsymbol{U}_d^T \hat{\boldsymbol{\mu}}_A \quad (8)$$

The same calculation is done for $\boldsymbol{\mu}_B$. Using these quantities we estimate $\boldsymbol{f} = \frac{\boldsymbol{\mu}_A - \boldsymbol{\mu}_B}{\|\boldsymbol{\mu}_A - \boldsymbol{\mu}_B\|}$. Then, we process an unknown point $\boldsymbol{p}$ by transforming it with $\boldsymbol{W}_D$, and projecting it on $\boldsymbol{f}$; both these steps are performed by the inner product $\boldsymbol{w} \cdot \boldsymbol{p}$, where:

$$\boldsymbol{w} = \boldsymbol{W}_D^T \boldsymbol{f} = q_d \boldsymbol{U}_d^T \boldsymbol{Q}_d^{-1} \boldsymbol{U}_d \boldsymbol{f} + \boldsymbol{f} - \boldsymbol{U}_d^T \boldsymbol{U}_d \boldsymbol{f} \quad (9)$$

Finally, given $\gamma$ as in Equation (4), $\boldsymbol{p}$ is assigned to class $A$ if $\boldsymbol{w} \cdot \boldsymbol{p} \geq \gamma$, to class $B$ otherwise. It is important to notice that this thresholding approach is robust to unbalanced classes as noticed in [52]. This is particularly suitable for ethnicity classification (see next Section).

### 5.2 Ethnicity Estimation

To estimate ethnicity we have to deal with an unbalanced multi-class classification problem. For this reason we have combined binary classifiers to obtain a multi-class classification approach. Precisely, we have employed a Decision Direct Acyclic Graph (`DDAG`, [47]) to combine `TIPCAC` classifiers. It is important to notice that the combination of binary classifiers usually guarantees better results with respect to native multi-class classification approaches. Furthermore, we have chosen to employ `DDAG+TIPCAC` since it is fast to evaluate and since `TIPCAC` maintains high accuracy levels also when dealing with high unbalanced classes as shown in [52].

A Rooted Direct Acyclic Graph (`DAG`) is a graph whose edges have an orientation, no cycles, and only one root node. A Rooted Binary `DAG` has nodes which have either 0 or 2 arcs leaving them. A `DDAG` is a method that combines the results of *one-against-one* classifiers to produce a multiclass classification. To this aim, considering a $N$-class problem, the `DDAG` is implemented using a rooted binary `DAG` with $K = N(N-1)/2$ internal nodes. Each node represents a classification model trained on two of the $K$ classes, and it produces a boolean output value ($\{0, 1\}$). The nodes are arranged in a binary tree with the single root node at the top, two nodes in the second layer and so on until the final layer of leaves. Considering each classifiers as a boolean function, to perform a classification the `DDAG` proceeds as follows: it starts at the root node and it evaluates

the boolean function; the node is then exited via the left edge, if the binary function is zero, or the right edge, if the binary function is one; the next nodes binary function is then evaluated; the membership class is the final leaf node reached through this process.

### 5.3 Age Estimation

Regression is the natural approach to solve age estimation, indeed most of the existing state-of-the art methods employ it at some stage. Regression, when enough data is available, allows to learn more general models, also for labels that are not given at training time. Moreover predicting age by regression avoids errors due to quantization of the age variable.

Considering the high dimensionality of our feature, we can obtain good performance with a regularized linear regressor. Linear regressors have several advantages, especially from an applicative point of view. First, memory foot print is reduced, requiring to memorize a single weight vector; second, avoiding kernels or deep learning based approach, the method is extremely efficient, requiring only the computation of a dot product among a face feature and the learned regressor.

We estimate a weight vector $\boldsymbol{w}$ and a bias $b$ in order to produce an age estimate given an image $\boldsymbol{I}$ and a feature function $\phi(\cdot)$:

$$\text{age}(\phi(\boldsymbol{I})) = \langle \boldsymbol{w}, \phi(\boldsymbol{I}) \rangle + b \tag{10}$$

We efficiently learn weights applying gradient based learning to a regularized least square problem:

$$\frac{1}{2}\lambda ||\boldsymbol{w}||^2 + \frac{1}{n}\sum_{i=1}^{N} \left(\langle \boldsymbol{w}, \phi(\boldsymbol{I}) \rangle + b - y_i\right)^2 \tag{11}$$

For large datasets stocastic gradient descent is accurate and very efficient [4]. We set $\lambda = 1/(C \cdot N))$, where $N$ are the training samples, and tune the parameter C by five fold cross-validation of MAE on the training set.

It is important to underline that in a specific experimental test on age estimation (see Section 6.4 paragraph a) we have employed the same methodology described in Section 5.2. We have employed this approach when the age labels are quantized in a set of age groups since, under this setting, a regression technique is excessively penalized.

### 5.4 Profiling on video sequences

To cope with the issues raised by real-world scenarios, in which a profiling algorithm should output a prediction after observing a set of possibly noisy and misaligned face patches, we introduce our approach for video sequences.

We propose a smart strategy to aggregate predictions based on an estimated quality, which we use as a weight. In the following we introduce an image quality measure and the final decision model to incorporate multiple frames into a single decision for a track extracted using the method in Section 4.4.

Consider a set of decisions $\mathcal{Y}_a : \{y_a(1) \dots y_a(T)\}$ for an attribute $a$ performed on samples of a track $\mathcal{T}$. We compute a measure of quality for each sample, using symmetry. Face quality estimation is usually performed measuring head pose, illumination and sharpness[17, 42]. Our feature is based on local gradient features, which are robust to illumination issues. Moreover, it is not clear what lighting conditions are best to perform face profiling. Head pose is the most important cue for face quality, considering the fact that the majority of the datasets on which profilers can be learned present mostly frontal images. Estimating the 3D pose in real-time is a challenging problem and performing a full 3D frontalization is expensive computationally.

Symmetry is a fast tool to verify the quality of a detected face. It has two main advantages: first, it is computed extremely fast by comparing original and flipped version of aligned faces. Second, it accounts for both the presence of partial facial occlusions and the yaw of a face with respect to the camera; indeed, even when alignment is applied, artifacts may appear when the detected face is at an angle, which is more than 20°off the camera center.

Let $I(t)$ be the imagery of an aligned face and $I_{lr}(t)$ its horizontally flipped version. We compute weights using the following:

$$w_t = 1 - \frac{|I(t) - I_{lr}(t)|}{S^2}$$

We further normalize weights taking into account the track length $w'_t = w_t/|\mathcal{T}|$. We implement two different strategies to make decision on track attributes aggregating the weighted decisions of every single frame. For categorical variables, we compute weighted histograms of attribute value counts. Being $b_i^a = \sum_{k=1}^{N} w_k$ the bin accounting for attribute $a$ having value $i$; decision on the track is taken as $\hat{y}_{\mathcal{T}} = \arg\max_i b_i$.

For age, which is a continuous value, we use the weighted median as a robust estimator to avoid the mean being deviated by few outliers. This phenomenon is already moderated by the use of weights, however we found out this strategy to be more robust to noise.
The weighted median of an ordered set of ages $\{a_1 \dots a_T\}$, if weights are normalized, as in our case, is found as the

element $a_m$ for which:

$$\sum_{i=1}^{m-1} w_i < \frac{1}{2} \text{ and } \sum_{i=m+1}^{|\mathcal{T}|} w_i \geq \frac{1}{2}.$$

In case $\sum_{i=1}^{m} w_i = \frac{1}{2}$ the median is found as $a_m = \frac{1}{2}(a_m + a_{m+l})$ where $m + l$ is the next profile with non zero weight.

## 6 Experimental Results

We tested our approach on five datasets the MORPH-II, Adience, Chalearn (LAP2015 and LAP2016) and the McGill Faces datasets.

MORPH-II[4] contains more than 55K facial images with different gender and ethnicity. In Tab. 6 the detailed statistics of gender and ethnicity are shown, whilst in Fig. 6 the age distribution is summarized.

Adience[5] is a challenging dataset collected from selfie images posted on Flickr. It is composed by 26,580 photos of 2,284 subjects. Exact age is not reported, rather, it is classified into 1 of 8 ranges.

Chalearn [16] is a competition to estimate apparent age from face imagery. In Looking At People (LAP) 2015 competition a dataset[6] of 5,000 images displaying a single individual where collected. In LAP2016[7] 8,000 images where collected. In both cases all images were annotated by a multiple annotators and average and standard deviation are released instead of ground truth age values.

McGill[8] was collected by DeMirkus *et al.* [12] which acquired videos from 60 subjects. The public release of the dataset accounts for 35 subjects. The dataset contains videos of subjects performing natural actions such as drinking from a cup, removing sunglasses, talking with arbitrary poses. The whole dataset accounts for more than 10K face images. Results on this dataset are especially interesting to test how our tracking based profiling improves over framewise estimation.

We ran a set of experiments to evaluate the effect of our system parameters. We used the challenging age estimation problem as a benchmark on MORPH-II. Considering the results reported in Sect. 6.3 and the accuracy/efficiency trade-off, we used 128 Gaussians as the codebook size, set the sampling stride to 6 and the scales to 4 and 8 for all experiments in every dataset.

[4] http://goo.gl/NKVCam

[5] https://www.openu.ac.il/home/hassner/Adience/data.html

[6] http://chalearnlap.cvc.uab.es/dataset/18/description/

[7] http://chalearnlap.cvc.uab.es/dataset/19/description/

[8] A video with showing the output of our algorithm on every subject in this dataset is available as supplementary material.

| System | FPS |
|---|---|
| Our Approach | 20 |
| Junyu Tech. | 15 |
| Zhuhau-Yisheng | 10 |
| MITRE | 27 |
| Tsinghua University | 11 |
| NEC | 19 |
| Cognitech | 5 |

Table 1: FPS of commercial systems reported in [44, 45]. Algorithm timing is referred to the full evaluation from pixel to prediction on gender and age prediction.

### 6.1 Timing

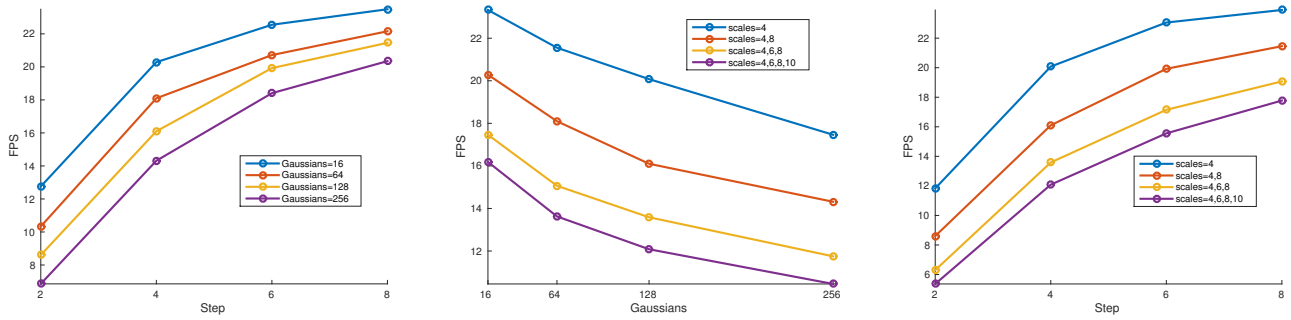
We run a set of benchmarks to evaluate the run time of our method using a i5-2467M 1.60GHz CPU. The system speed is mostly affected by the density of feature sampling both in scale and size as can be seen in Figs. 4a and 4c since the sampling step quadratically affects the amount of features extracted. Moreover, when using larger codebooks the amount of Gaussians has two effects on the computational cost. First, Fisher Vector embeddings require more time since the derivatives to compute are depend linearly with the number of Gaussians. Second, a larger codebook increases the feature dimensionality and consequently the classification time, although the classifier time is negligible with respect to the feature extraction and embedding cost.

In Tab. 1 we have reported the FPS of some commercial systems presented in [44, 45]. On both technical reports Ngan *et al.* report the same plot for timing, meaning that commercial algorithms are evaluated on the full demographic profiling task, i.e. predicting age and gender. We report our timing as the time needed to detect, align a face and predict ethnicity, gender and age. We are therefore solving all three tasks at once, while the algorithms tested in these reports, to the best of our knowledge, only address gender and age estimation.

The best commercial frameworks obtain a performance comparably to our approach, although it has to be noted that the system that is used to test this algorithms is a 6-cores Intel Xeon Processor X5690, which is by far more powerful than our 1.6Ghz i5-2467M processor. Moreover their performance figures are measured using multi-threading. This analysis confirms that our method has state-of-the art performance.

### 6.2 Gender Estimation

We evaluate the gender estimation accuracy using mean per class accuracy in order to deal with the imbalance of some datasets.




(a) (b) (c)

Fig. 4: Frame rate of the proposed processing pipeline for different dictionary size, sampling in space and scale. We set scales=4,8 in (a) step=6 (b) and Gaussians=128 (c). Face detection and alignment is included.

| Method | Gender | time (ms) | GPU |
|---|---|---|---|
| Our | 80.1 | **50** | n |
| Levi *et al.* [37] | **85.9** | 200 | y |
| Eidinger *et al.* [15] | 77.8 | ? | n |
| Hassner *et al.* [29] | 79.3 | ? | n |

Table 2: Gender estimation accuracy on the Adience dataset compared with the state of the art. We report, when available face profiling timing and if methods employ a high-end GPU to produce such timing.

| Method | Accuracy |
|---|---|
| KCCA [23] | 98.5 |
| KPLS [22] | 98.4 |
| **Our** | 98.2 |
| 3-Step[22] | 98.1 |
| rCCA[23] | 97.6 |
| PLS [22] | 97.4 |
| CCA [23] | 95.2 |

Table 3: Gender estimation accuracy on the MORPH-II dataset compared with the state of the art.

#### *6.2.1 Adience*

On the challenging Adience dataset our method accuracy is comparable with those achieved by previously published methods not using deep learning. On this binary task there are a sufficient amount of samples to obtain a better predictor using CNN features. Although our approach is still 4 times faster running on a single threaded CPU (instead of multi-GPU cores) as shown in Tab. 2.

#### *6.2.2 Morph-II*

In Tab. 3 we report our result on Morph-II. In this dataset we have a very high accuracy in gender estimation with a mean per class accuracy over 98%. However, this dataset is easier than Adience and McGill, having little or no variations in pose and illumination.

#### *6.2.3 McGill*

We report results on video sequences taken from McGill dataset. Our method exploits a tracker thus allowing to improve attribute prediction when performed on a sequence. Tab. 4 for each accuracy result reports if the temporal information has been used (Time column) and what is the granularity of the evaluation (Evaluation column).

In Tab. 4 we compare the gender prediction obtained using the method described in Sect. 5.4 in the two settings proposed by DeMirkus *et al.* [12].

In both settings, the accuracy is evaluated on a per-frame basis. In the first setting, the temporal information is discarded and each frame is treated as an independent image. In the second setting a tracker is used and the temporal information is exploited. The first setting evaluates the quality of the classifier and the model; our method outperforms theirs by 5%. The latter setting is a more realistic one, showing the performance obtainable in a real world application; in this scenario we again obtain a 5% improvement over the method proposed by DeMirkus *et al.* [12].

Finally, we report two results not directly comparable to [12]. Instead of evaluating a frame by frame correctness we evaluate on a per-track and per-sequence basis. We found this settings more realistic since in a typical non-cooperative scenario a person will walk towards the profiling camera and the system will capture one or more track of the subject. Our method applied on the whole track has a slight improvement. Finally, if all tracks predictions are combined through majority

| Time | Evaluation | DeMirkus | Our Method |
|---|---|---|---|
| - | Frame | 0.6960 | **0.7451** |
| ✓ | Frame | 0.8524 | **0.8959** |
| ✓ | Track | - | **0.9104** |
| ✓ | Sequence | - | **0.9143** |

Table 4: Accuracy on gender estimation. We report accuracy computed averaging profiling output on Frames, Tracks and Sequences.

| Time | Evaluation | Our Method |
|---|---|---|
| - | Frame | 0.6904 |
| ✓ | Frame | 0.7402 |
| ✓ | Track | 0.7424 |
| ✓ | Sequence | 0.7429 |

Table 5: Accuracy on ethnicity estimation. We report accuracy computed averaging profiling output on Frames, Tracks and Sequences.

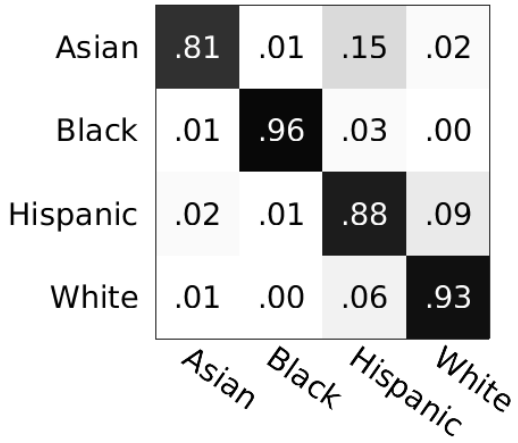



Fig. 5: Confusion matrix for ethnicity estimation on MORPH-II.

voting for each sequence we obtain an even higher performance. To the best of our knowledge, this results are the state-of-the art on this dataset.

### 6.3 Ethnicity Estimation

Predicting people ethnicity has been addressed by few works. In this section, we report a comparison with the results achieved in [21] on the MORPH-II dataset. The dataset is highly unbalanced having more than 75% of black subjects. In Fig. 5 the confusion matrix for ethnicity estimation is shown.

We have obtained a mean per class accuracy of 89.61% while Guo *et al.* have reported 82.3%. Our `DAG-TIPCAC` classifier is able to deal better with this highly unbalanced dataset. The high bias is due to the fact that this data contains few Hispanic and Asian samples and this has an evident effect also on the lower accuracies obtained on these two classes.

Finally, to address the real world setting of profiling from video sequences we labeled with the four races of MORPH-II the subjects in the McGill dataset and reported results in Tab. 5. Ethnicity estimation is harder than gender estimation, nonetheless integrating information over time improves by 5% points. Accuracy estimated on tracks and whole sequences is slightly superior.

### 6.4 Age Estimation

Age estimation can be defined as the prediction of the exact age or as the prediction of an age class. Some dataset do not provide the exact age value therefore we evaluated our method using a classifier, reporting exact age estimation accuracy and one-off age estimation error. In case one-off error is employed, the age class is considered correct, even if the predicted age range is one of the two adjacent ones.

When a sufficient amount of data is available we have assessed the quality of our method using the Mean Absolute Error or MAE $= \frac{1}{N}\sum_{i=1}^{N}|\text{age}(\phi(\boldsymbol{I})) - y_i|$. In certain evaluation protocols the apparent age is used, collecting average $\mu_i$ and standard deviation $\sigma$ of age for each picture $i$ from a set of annotators. In such cases standardized error is used:

$$\epsilon = 1 - \exp\left(-\frac{(\text{age}(\phi(\boldsymbol{I})) - \mu_i)^2}{2\sigma^2}\right) \tag{12}$$

#### *6.4.1 MORPH-II*

On this dataset we perform experiments using three different setups in order to be comparable with recently published results. The outcome of experiments in all three settings are summarized in Table8

We replicated the experiments in [9,8] using a set of 5,492 pictures of people of Caucasian ancestry, averaging the MAE over 30 runs. We used the same pictures used by the authors.

Guo *et al.* perform experiments on a larger set of images with a slightly more complex procedure [24,20, 22]. Considering the whole set $\mathcal{W}$, a set $\mathcal{S} \subset \mathcal{W}$ of $\sim$ 21,000 images is formed from black and white individuals keeping all the women and adding male subjects mantaining the ratio between males and females to 3:1. Set $\mathcal{S}$ is furtherly partitioned in two disjoint sets $\mathcal{S}_1, \mathcal{S}_2$ so that identities of people in one set are not allowed in

| Race | Female | Male | Female and Male |
|---|---|---|---|
| Black | 5,757 | 36,803 | 42,560 |
| White | 2,601 | 7,999 | 10,600 |
| Hispanic | 100 | 1,651 | 1,751 |
| Asia | 13 | 146 | 159 |
| India | 14 | 43 | 57 |
| Other | 2 | 3 | 5 |
| Total | 8,487 | 46,645 | 55,132 |

Table 6: MORPH-II dataset gender and ethnicity statistics.

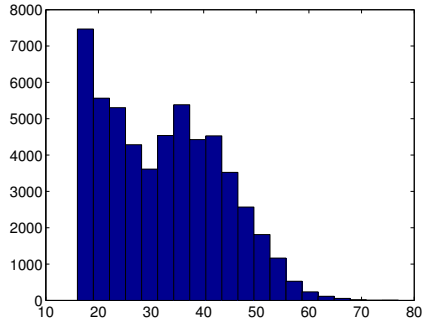



Fig. 6: MORPH-II age distribution.

the other and vice versa. We trained our system on both subsets $\mathcal{S}_i$ and reported the average MAE computed on $\mathcal{W} \setminus \mathcal{S}_i$ for $i = 1, 2$.

Finally, we use the setup proposed in [19]. The dataset is split using 80% of identities for training and the remaining data for testing, running a 10-fold cross-validation. Even if we do not stratify the sampling for gender and ethnicity, we could check empirically that random identity sampling guarantees to keep age, gender and ethnicity distributions on training and testing sets.

Using MAE as a metric to assess performance can be prone to misinterpretations. We therefore show a more insightful analysis of our regression based age estimation on MORPH-II showing the Cumulative Score curve (CS). The cumulative score curve $CS(t)$ is defined as the amount of samples for which the estimation error is lower than a threshold t in years. Fig. 7 shows that 75% of the predictions have an error lower than 5 years.

We first report an analysis of how error varies depending on parameters affecting the feature extraction process. Table 7 shows MAE variation depending on SIFT density, scales and codebook Gaussians using the setting of [19]. Interestingly enough, the amount of scales does not affect the error, while a too wider sampling or too few Gaussians are detrimental to the accuracy. Setting the SIFT sampling step to 4 instead of 2 does not change the error while, increasing the extraction and coding efficiency quadratically. Using 128 Gaussians instead of 256 represents a good trade-off in order not to sacrifice efficiency, as also results from Fig. 4.

| Scales | Sampling | MAE |
|---|---|---|
| 4,6,8,10 | 2 | 3.7 |
| 4,8 | 2 | 3.7 |
| 4,8 | 6 | 3.7 |
| 4,8 | 8 | 4.0 |

(a)

| Gaussians | MAE |
|---|---|
| 16 | 4.2 |
| 64 | 3.8 |
| 128 | 3.7 |
| 256 | 3.6 |

(b)

Table 7: Mean absolute error varying sampling step, scales and Gaussians. We used 128 Gaussians in (a) and step=6 and scales=4,8 in (b). The algorithm is mostly affected by the sampling step.

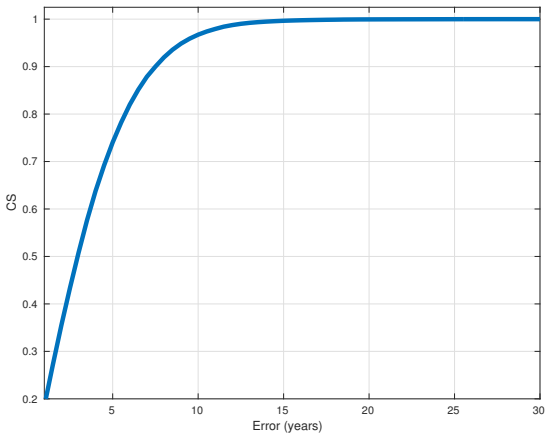



Fig. 7: Cumulative Score curve on MORPH-II. The curve $CS(t)$ measure the percentage of samples have an age predicted with an error lower than $t$.

Results presented in Tab. 8 are obtained with three very different setups. The one proposed by Guo *et al.* is the easiest, employing a single ethnicity[24, 20, 22]. Multiple ethnicities are taken into consideration in the second and third setups. Chang *et al.* use only black and white individuals[9, 8] while Geng *et al.* use the whole dataset [19]. We beat all previously reported result in every setting, this shows that our age estimator works in cross-racial and cross-gender settings, without any specific precautions. Recently methods using ensemble of large CNN models reported results in certain cases improving MAE with respect to our results[26, 13], with a MAE of 4.0 and 2.6 respectively. We did not report the results in Tab. 8 since the authors did not clarify what is the data split they have used and results are not directly comparable.

*Cross-Domain Evaluation* It is interesting to understand how much age depends on ethnicity and gender. In this setting, named cross-domain, we train on a single demographic, e.g. Black Females, and test on the remaining ones. It can be seen that highest MAE is given when both ethnicity and gender are swapped between train and test set. Interestingly, all methods perform

| Approach | Features | Classifier | MAE [9,8] | MAE[24,20,22] | MAE[19] |
|---|---|---|---|---|---|
| Our approach | SIFT+FV | L2L2 Regression | 3.8 | 4.0 | 3.7 |
| Geng *et al.* [19] | AAM,BIF | CPDNN | - | - | 4.9 |
| Geng *et al.* [19] | AAM,BIF | IIS-LLD | - | - | 5.7 |
| Huerta *et al.* [32] | SURF/HOG | CCA | - | - | 4.2 |
| Guo *et al.* [22] | Holistic BIF | Kernel PLS | - | 4.2 | - |
| Guo *et al.* [20] | Holistic BIF | 3-Step | - | 4.5 | - |
| Guo *et al.* [24] | Holistic BIF | Linear SVM | - | 5.1 | - |
| Chang *et al.* [9] | AAM | Ordinal Hyperplane Ranker | 6.1 | - | |
| Chang *et al.* [8] | AAM | Ranking SVM | 6.5 | - | - |

Table 8: Mean Absolute Error (MAE) in years compared with recently published methods. Our method obtains results comparable with the state-of-the-art with a very low-weight processing pipeline.

| Train | Test | Guo *et al.* [20] | CpDA [25] | Ours |
|---|---|---|---|---|
| WF | BF | 8.67 | 6.54 | **6.32** |
| | WM | 7.72 | 5.57 | **4.94** |
| | BM | 10.62 | 7.54 | **7.45** |
| BF | WF | 9.15 | 6.41 | **5.16** |
| | BM | 8.40 | 6.13 | **5.15** |
| | WM | 8.79 | 6.23 | **6.06** |
| WM | BM | 7.05 | 5.35 | **4.38** |
| | WF | 9.13 | 6.70 | **6.66** |
| | BF | 9.54 | 7.67 | **6.50** |
| BM | WM | 6.86 | 5.10 | **4.69** |
| | BF | 10.58 | 7.73 | **4.87** |
| | WF | 12.81 | 8.73 | **5.19** |

Table 9: Mean Absolute Error(years), in cross-domain setting, using the split proposed by [20], using only Black(B) and White(W), Males(M) and Females(F)

better in cross-ethnicity setting than in cross-gender setting.

### *6.4.2 Adience*

We employ the five fold cross-validation using the folds provided by the authors. In Tab. 10 we report age estimation performance together with the timing of our approach. Our method achieves a good trade-off between computational cost and age estimation error. Consider that Levi *et al.* train a deep neural network on face images. Deep convolutional networks need a sufficient amount of data not to overfit and learn strong classifiers. In the case of gender their method outperform ours, although in the age class estimation we perform better since our technique can leverage also lower amount of data. Finally, comparing the timing our single-threaded CPU based method is four times faster than the one proposed by Levi which requires a high end GPU. Recent results, exploiting larger deep architectures are able to outperform our method, nonetheless they are extremely demanding in terms of hardware both to attain real-time performance and to be able to run larger models such as VGG-16.

| Method | Age (exact) | Age (1-off) | time ms | GPU |
|---|---|---|---|---|
| Duan *et al.* [13] | **66.5** | - | ? | y |
| Hou *et al.* [30] | 64.0 | 96.6 | ? | y |
| Gürpinar *et al.* [26] | 51.3 | - | ? | y |
| Our | 50.8 | 84.3 | **50** | n |
| Levi*et al.* [37] | 49.5 | 84.6 | 200 | y |
| Eidinger *et al.* [15] | 45.1 | 79.5 | ? | n |

Table 10: Age class estimation accuracy of our method compared with state of the art techniques. We report, when available face profiling timing and if methods employ a high-end GPU to produce such timing.

### *6.4.3 LAP2015 and LAP2016*

In this section we test our approach on the novel challenging datasets LAP2015 and LAP2016. To run a fair comparison we report the $\epsilon$-error and the time required to extract features and classify an image. Most of the competing approach are based on ensembles of fine-tuned VGG-16 networks. VGG-16 has 150M parameters, can not be run on GPUs with less than 6Gb of RAM even for small batch sizes and requires 31 G-Ops as reported in [7]. To make a fair comparison in terms of required computational power we compare timings obtained running on CPU the base architecture reported in [16], also accounting for augmentation and ensembling as in the case of the leading submission for LAP2016 by OrangeLabs[6]. In case the CNN architecture was not available or described in detail we did not report the timing. Detection and alignment times are kept out of the timing for competing methods. In our case we report the full pipeline cost. Note that most of these methods are also using Deep Convolutional Neural Network for detection and alignment.

On LAP2015 our approach is just below human performance, but much faster, from 2X to 800X in case of the best performing method which uses an ensemble of 20 CNNs. We have a similar performance on LAP2016 and also in this case we are from 800X to 3X faster then competing approaches.

| Approach | eps score | time (s) |
|---|---|---|
| CVL ETHZ [49] | 0.264975 | 40.0 |
| ICT-VIPL [38] | 0.270685 | 0.10 |
| AgeSeer | 0.287266 | 0.50 |
| WVU CVL [63] | 0.294835 | 0.10 |
| SEU-NJU [60] | 0.305763 | 0.50 |
| *human reference* | *0.34* | - |
| Ours | 0.354066 | **0.05** |
| UMD | 0.373352 | |
| Enjuto | 0.374390 | - |
| Sungbin Choi | 0.420554 | - |
| Lab219A | 0.499181 | - |
| Bogazici | 0.524055 | - |
| Notts CVLab | 0.594248 | - |

Table 11: Age estimation results on LAP15 in terms of $\epsilon$-error and classification time for a single sample using the same CPU.

| Approach | eps score | time (s) |
|---|---|---|
| OrangeLabs[6] | 0.2411 | 44.0 |
| palm seu[33] | 0.3214 | 2.00 |
| cmp+ETH[58] | 0.3361 | 0.50 |
| WYU CVL | 0.3405 | - |
| ITU SiMiT[40] | 0.3668 | 1.50 |
| Ours | 0.3684 | **0.05** |
| Bogazici | 0.3740 | 0.50 |
| MIPAL SNU | 0.4569 | 0.15 |
| DeepAge | 0.4573 | 0.50 |

Table 12: Age estimation results on LAP16 in terms of $\epsilon$-error and classification time for a single sample using the same CPU.

### *6.4.4 McGill*

In this experiment we report results on video for age estimation. Considering that real age value were not available we asked a set of 40 human annotators to examine a single frame from each video and express an age for each subject. We handpicked frames to select neutral expressions, small blur and a frontal pose to avoid any bias. Annotators had an average standard deviation of 4.47 years. We report MAE with respect to the estimated mean. As it is shown in Tab. 13 the behavior is consistent with Tab. 4 for gender, showing a clear improvement from our tracking algorithm. Note that MAE figures are higher in this dataset with respect to MORPH-II due to pose and expression variation as well as blur and varying lighting conditions.

| Time | Evaluation | Our Method |
|---|---|---|
| - | Frame | 8.5902 |
| ✓ | Frame | 7.1687 |
| ✓ | Track | 6.9439 |
| ✓ | Sequence | 6.2413 |

Table 13: MAE on age estimation in video. We report MAE computed averaging profiling output on Frames, Tracks and Sequences. Our method exploiting tracking allows a reduction of more than 2 years of error.

## 7 Conclusions

We have proposed a demographic profiling algorithm that estimates age, gender and ethnicity from face imagery. Our method is carefully designed to attain real-time performance. We hence showed how an efficient image processing pipeline, combined with the latest generation hand-crafted features an classifiers can deliver the sought result. Our system can run on a single thread at approximately 20 FPS on a i5-2467M processor, clocking a 1.6GHz. Our approach can indeed be speed up by incorporating CUDA implementation of its core components, e.g. using[3] for feature extraction and CuBLAS[1] for classifiers and other linear algebra operations. Low power single board GPU systems are nowadays available[2] but they are not comparable in terms of cost with respect to their CPU-only counterparts.

To assess the quality of our framework we have tested our approach on large datasets of people images with different age, gender and ethnicity. Our method results compared with those achieved by other recently published approaches confirm the efficiency and the effectiveness of the proposed framework.

**Acknowledgments.**
Lorenzo Seidenari is partially supported by "THE SOCIAL MUSEUM AND SMART TOURISM", MIUR project no. CTN01_00034_23154_SMST.

## References

1. Nvidia cublas library for fast gpu-accelerated implementation of the standard basic linear algebra subroutines (blas). https://developer.nvidia.com/cublas. Accessed: 2018-09-18.
2. Nvidia embedded gpu boards. https://developer.nvidia.com/embedded/develop/hardware. Accessed: 2018-09-18.
3. K Aniruddha Acharya, R Venkatesh Babu, and Sathish S Vadhiyar. A real-time implementation of sift using gpu. *Journal of Real-Time Image Processing*, 14(2):267–277, 2018.
4. Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid. Good practice in large-scale learning for image classification. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(3):507–520, March 2014.
5. Lus A. Alexandre. Gender recognition: A multiscale decision fusion approach. *Pattern Recognition Letters*, 31(11):1422–1427, 2010.
6. Grigory Antipov, Moez Baccouche, Sid-Ahmed Berrani, and Jean-Luc Dugelay. Apparent age estimation from face images combining general and children-specialized deep learning models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 96–104, 2016.

7. Alfredo Canziani, Adam Paszke, and Eugenio Culurciello. An analysis of deep neural network models for practical applications. *arXiv preprint arXiv:1605.07678*, 2016.
8. Kuang-Yu Chang, Chu-Song Chen, and Yi-Ping Hung. A ranking approach for human ages estimation based on face images. In *Proc. of ICPR*, pages 3396–3399, Aug 2010.
9. Kuang-Yu Chang, Chu-Song Chen, and Yi-Ping Hung. Ordinal hyperplanes ranker with cost sensitivities for age estimation. In *Proc. of CVPR*, pages 585–592, June 2011.
10. Marcos Vinicius Mussel Cirne and Helio Pedrini. Gender recognition from face images using a geometric descriptor. In *Systems, Man, and Cybernetics (SMC), 2017 IEEE International Conference on*, pages 2006–2011. IEEE, 2017.
11. T.F. Cootes, G.J. Edwards, and C.J. Taylor. Active appearance models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(6):681–685, Jun 2001.
12. Meltem Demirkus, Doina Precup, James Clark, and Tal Arbel. Hierarchical spatio-temporal probabilistic graphical model with multiple feature fusion for estimating binary facial attribute classes in real-world face videos. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2015.
13. Mingxing Duan, Kenli Li, and Keqin Li. An ensemble cnn2elm for age estimation. *IEEE Transactions on Information Forensics and Security*, 13(3):758–772, 2018.
14. Mingxing Duan, Kenli Li, Canqun Yang, and Keqin Li. A hybrid deep learning cnn–elm for age and gender classification. *Neurocomputing*, 275:448–461, 2018.
15. Eran Eidinger, Roee Enbar, and Tal Hassner. Age and gender estimation of unfiltered faces. *Information Forensics and Security, IEEE Transactions on*, 9(12):2170–2179, 2014.
16. Sergio Escalera, Mercedes Torres Torres, Brais Martinez, Xavier Baró, Hugo Jair Escalante, Isabelle Guyon, Georgios Tzimiropoulos, Ciprian Corneou, Marc Oliu, Mohammad Ali Bagheri, et al. Chalearn looking at people and faces of the world: Face analysis workshop and challenge 2016. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8, 2016.
17. A. Fourney and R. Laganiere. Constructing face image logs that are both complete and concise. In *Proc. of CRV*, pages 488–494, 2007.
18. Yun Fu and Thomas S. Huang. Human age estimation with regression on discriminative aging manifold. *Multimedia, IEEE Transactions on*, 10(4):578–584, 2008.
19. Xin Geng, Chao Yin, and Zhi-Hua Zhou. Facial age estimation by learning from label distributions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(10):2401–2412, Oct 2013.
20. Guodong Guo and Guowang Mu. Human age estimation: What is the influence across race and gender? In *Proc. of CVPRW*, pages 71–78, June 2010.
21. Guodong Guo and Guowang Mu. A study of large-scale ethnicity estimation with gender and age variations. In *Proc. of CVPRW*, pages 79–86. IEEE, 2010.
22. Guodong Guo and Guowang Mu. Simultaneous dimensionality reduction and human age estimation via kernel partial least squares regression. In *Proc. of CVPR*, pages 657–664, June 2011.
23. Guodong Guo and Guowang Mu. Joint estimation of age, gender and ethnicity: Cca vs. pls. In *Proc. of FG*, pages 1–6. IEEE, 2013.
24. Guodong Guo, Guowang Mu, Yun Fu, and T.S. Huang. Human age estimation using bio-inspired features. In *Proc. of CVPR*, pages 112–119, June 2009.
25. Guodong Guo and Chao Zhang. A study on cross-population age estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4257–4263, 2014.
26. Furkan Gurpinar, Heysem Kaya, Hamdi Dibeklioglu, and Ali Salah. Kernel elm and cnn based facial age estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 80–86, 2016.
27. Hu Han, Charles Otto, and Anil K. Jain. Age estimation from face images: Human vs. machine performance. In *Proc. of ICB*, 2013.
28. Per C. Hansen. The truncated SVD as a method for regularization. *BIT Numerical Mathematics*, 27(4):534–553, December 1987.
29. Tal Hassner, Shai Harel, Eran Paz, and Roee Enbar. Effective face frontalization in unconstrained images. In *Proc. of CVPR*, June 2015.
30. Le Hou, Chen-Ping Yu, and Dimitris Samaras. Squared earth movers distance loss for training deep neural networks on ordered-classes. In *Proc. of NIPS*, 2017.
31. Guang-Bin Huang, Lei Chen, Chee Kheong Siew, et al. Universal approximation using incremental constructive feedforward networks with random hidden nodes. *IEEE Trans. Neural Networks*, 17(4):879–892, 2006.
32. Ivan Huerta, Carles Fernández, and Andrea Prati. Facial age estimation through the fusion of texture and local appearance descriptors. In *European Conference on Computer Vision*, pages 667–681. Springer, 2014.
33. Zengwei Huo, Xu Yang, Chao Xing, Ying Zhou, Peng Hou, Jiaqi Lv, and Xin Geng. Deep age distribution learning for apparent age estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 17–24, 2016.
34. Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proc. of CVPR*, 2014.
35. D. E. King. Max-Margin Object Detection. *ArXiv e-prints*, January 2015.
36. Sebastian Lapuschkin, Alexander Binder, Klaus-Robert Müller, and Wojciech Samek. Understanding and comparing deep neural networks for age and gender classification. *arXiv preprint arXiv:1708.07689*, 2017.
37. Gil Levi and Tal Hassner. Age and gender classification using convolutional neural networks. In *Proc. of CVPRW*, June 2015.
38. Xin Liu, Shaoxin Li, Meina Kan, Jie Zhang, Shuzhe Wu, Wenxian Liu, Hu Han, Shiguang Shan, and Xilin Chen. Agenet: Deeply learned regressor and classifier for robust apparent age estimation. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*, December 2015.
39. David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal on Computer Vision*, 60(2):91–110, November 2004.
40. Refik Can Malli, Mehmet Aygün, and Hazim Kemal Ekenel. Apparent age estimation using ensemble of deep learning models. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2016 IEEE Conference on*, pages 714–721. IEEE, 2016.
41. Ethan Meyers and Lior Wolf. Using biologically inspired features for face processing. *International Journal of Computer Vision*, 76(1):93–104, 2008.

42. Kamal Nasrollahi and Thomas B Moeslund. Face quality assessment system in video sequences. In *Biometrics and Identity Management*, pages 10–18. Springer, 2008.
43. Choon Boon Ng, Yong Haur Tay, and Bok-Min Goi. Vision-based human gender recognition: A survey. *CoRR*, abs/1204.1611, 2012.
44. Mei Ngan and Patrick Grother. Face recognition vendor test (frvt) performance of automated age estimation algorithms. *NIST Interagency Report*, 7995, 2014.
45. Mei Ngan and Patrick Grother. Face recognition vendor test (frvt) performance of automated gender classification algorithms. In *Technical Report NIST IR 8052*. National Institute of Standards and Technology, 2015.
46. Bingbing Ni, Zheng Song, and Shuicheng Yan. Web image and video mining towards universal and robust age estimator. *Multimedia, IEEE Transactions on*, 13(6):1217–1229, 2011.
47. John C. Platt, Nello Cristianini, and John Shawe-taylor. Large margin DAGs for multiclass classification. In *Proc. of NIPS*, pages 547–553. MIT Press, 2000.
48. Maximilian Riesenhuber and Tomaso Poggio. Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2(11):1019–1025, 1999.
49. Rasmus Rothe, Radu Timofte, and Luc Van Gool. Dex: Deep expectation of apparent age from a single image. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*, December 2015.
50. A. Rozza, G. Lombardi, M. Rosa, and E. Casiraghi. O-ipcac and its application to eeg classification. *Proc. of WAPA*, pages 4–11, 2010.
51. Alessandro Rozza, Gabriele Lombardi, and Elena Casiraghi. Novel ipca-based classifiers and their application to spam filtering. *Proc. of ISDA*, pages 797–802, 2009.
52. Alessandro Rozza, Gabriele Lombardi, Elena Casiraghi, and Paola Campadelli. Novel fisher discriminant classifiers. *Pattern Recognition*, 45(10):3725–3737, October 2012.
53. Lorenzo Seidenari, Alessandro Rozza, and Alberto Del Bimbo. Real-time age estimation from face imagery using fisher vectors. In *Proc. of ICIAP*, 2015.
54. Lorenzo Seidenari, Giuseppe Serra, Andrew D. Badanov, and Alberto Del Bimbo. Local pyramidal descriptors for image recognition. *Pattern Analisys and Machine Intelligence, IEEE Transactions on*, 2013.
55. Du Shan and Ward R. Wavelet-based illumination normalization for face recognition. In *Proc. of ICPR*, 2005.
56. Karen Simonyan, Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. Fisher Vector Faces in the Wild. In *Proc. of British Machine Vision Conference (BMVC)*, 2013.
57. Jorge Snchez, Florent Perronnin, Thomas Mensink, and Jakob Verbeek. Image classification with the fisher vector: Theory and practice. *International Journal of Computer Vision*, 105(3):222–245, 2013.
58. Michal Uricár, Radu Timofte, Rasmus Rothe, Jiri Matas, and Luc Van Gool. Structured output svm prediction of apparent age, gender and smile from deep features. In *Proceedings CVPRW 2016*, pages 25–33, 2016.
59. Jian-Gang Wang, Jun Li, Wei-Yun Yau, and Eric Sung. Boosting dense SIFT descriptors and shape contexts of face images for gender recognition. In *Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE Workshop on*, pages 96–102, 2010.
60. Xu Yang, Bin-Bin Gao, Chao Xing, Zeng-Wei Huo, Xiu-Shen Wei, Ying Zhou, Jianxin Wu, and Xin Geng. Deep label distribution learning for apparent age estimation. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*, December 2015.
61. Ke Zhang, Ce Gao, Liru Guo, Miao Sun, Xingfang Yuan, Tony X Han, Zhenbing Zhao, and Baogang Li. Age group and gender estimation in the wild with deep ror architecture. *IEEE Access*, 5:22492–22503, 2017.
62. Ke Zhang, Miao Sun, Xu Han, Xingfang Yuan, Liru Guo, and Tao Liu. Residual networks of residual networks: Multilevel residual networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 2017.
63. Yu Zhu, Yan Li, Guowang Mu, and Guodong Guo. A study on apparent age estimation. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*, December 2015.