Received: 24 January 2021 | Revised: 3 May 2021 | Accepted: 4 May 2021

DOI: 10.1049/bme2.12046

IET Biometrics

The Institution of Engineering and Technology WILEY

ORIGINAL RESEARCH PAPER

Facial masks and soft-biometrics: Leveraging face recognition CNNs for age and gender prediction on mobile ocular images

Fernando Alonso-Fernandez[1] | **Kevin Hernandez-Diaz**[1] | **Silvia Ramis**[2] | **Francisco J. Perales**[2] | **Josef Bigun**[1]

[1]School of Information Technology, Halmstad University, Sweden

[2]Computer Graphics and Vision and AI Group, University of Balearic Islands, Spain

Correspondence

Fernando Alonso-Fernandez, School of Information Technology, Halmstad University, Sweden.
Email: feralo@hh.se

Funding information

University of Balearic Islands; Ministerio de EconomÃ-a y Competitividad, Grant/Award Number: PERGAMEX RTI2018-096986-B-C31; Ministerio de Ciencia e InnovaciÃ³n, Grant/Award Number: PID2019-104829RA-I00 / AEI / 10.13039/501100011033; VetenskapsrÃ¥det, Grant/Award Number: 2016-03497

Abstract

We address the use of selfie ocular images captured with smartphones to estimate age and gender. Partial face occlusion has become an issue due to the mandatory use of face masks. Also, the use of mobile devices has exploded, with the pandemic further accelerating the migration to digital services. However, state-of-the-art solutions in related tasks such as identity or expression recognition employ large Convolutional Neural Networks, whose use in mobile devices is infeasible due to hardware limitations and size restrictions of downloadable applications. To counteract this, we adapt two existing lightweight CNNs proposed in the context of the ImageNet Challenge, and two additional architectures proposed for mobile face recognition. Since datasets for soft-biometrics prediction using selfie images are limited, we counteract over-fitting by using networks pre-trained on ImageNet. Furthermore, some networks are further pre-trained for face recognition, for which very large training databases are available. Since both tasks employ similar input data, we hypothesise that such strategy can be beneficial for soft-biometrics estimation. A comprehensive study of the effects of different pre-training over the employed architectures is carried out, showing that, in most cases, a better accuracy is obtained after the networks have been fine-tuned for face recognition.

1 | INTRODUCTION

Recent research has explored the automatic extraction of information such as gender, age, ethnicity, etc. of an individual, known as soft-biometrics [1]. It can be deduced from biometric data like face photos, voice, gait, hand or body images, etc. One of the most natural ways is face analysis [2], but given the use of masks due to the COVID-19 pandemic, the face appears occluded even in cooperative settings, leaving the ocular region as the only visible part. In recent years, the ocular region has gained attention as a stand-alone modality for a variety of tasks, including person recognition [3], soft-biometrics estimation [4], or liveness detection [5]. Accordingly, this work is concerned with the challenge of estimating soft-biometrics when only the ocular region is available. Additionally, we are interested in mobile environments [6]. The pandemic has accelerated the migration to the digital domain, converting mobiles in data hubs used for all type of transactions [7]. In such context, selfie images are increasingly used in a variety of applications, so they enjoy huge popularity and acceptability [8]. Social networks or photo retouching are typical examples, but selfies are becoming common for authentication in online banking or payment services too.

Soft-biometrics information may not allow accurate person recognition, but in unconstrained scenarios where *hard* biometric traits (like face or iris) may suffer from degradation, it has been shown to improve the performance of the primary system [9]. If a sufficient number of characteristics are available, it might be even possible to carry out recognition with just soft-biometrics [10]. Such information has other diverse practical applications as well [1]. One example is targeted advertising, where customised products or services can be offered if age, gender or other characteristics of the customer are automatically inferred. In a similar vein, Human-Computer Interaction (HCI) can be greatly improved by knowing the particularities of the person who is interacting with the system.

In biometrics identification, search across large databases can be facilitated by filtering subjects with the same characteristics. On the one hand, it reduces the amount of comparisons, since only a portion of the database would be searched. On the other hand, it also allows to attain a better accuracy, since the errors of identification systems increases in proportion to the amount of comparisons [11]. Similarly, searches can be facilitated while looking for specific individuals in images or videos [12]. The complexity can be reduced enormously by searching or tracking persons only fulfiling certain semantic attributes (e.g. a young male with beard), filtering out those that are sufficiently distinct [13]. Another important fields of application are access control to products or services based on age (such as gambling, casinos, or games) and child pornography detection. The rapid growth of image and video collections due to high-bandwidth Internet and cheap storage is being accompanied by the necessity of efficient identification of child pornography, often within very large repositories of hundreds of thousands or millions of images [14].

Soft-biometrics using RGB ocular images captured by front cameras of smartphones (selfies) is a relatively new problem [15], with very few works [16–20]. Selfie images usually contain degradations like blur, uneven light and background, variable pose, poor resolution, etc. due to unconstrained environments and mobile operation. In addition, front cameras usually have lower quality in comparison to back-facing ones. In such conditions, soft-biometric attributes like gender or age may be extracted more reliably than features from primary biometric traits such as face or iris [21]. It may not even be necessary to look actively to the camera, so after initial authentication with a primary modality, the user may be continuously authenticated via soft-biometrics without active cooperation [22]. Transparent authentication is possible with other smartphone sensors as well, such as keystroke dynamics [23] or readings from the accelerometer or gyroscope [24]. Solutions to counteract the lack of resolution in primary modalities have been proposed too, such as super-resolution [25], so they are usable even at low resolution. However, the techniques in use are sensitive to acquisition conditions, degrading quickly with non-frontal view, illumination or expression changes. They also rely on a precise image alignment, which is an issue in low resolution, where blurring creates ambiguities for proper localization of facial landmarks or iris boundaries.

Another issue has to do with the limited resources of mobile devices. Recent developments in computer vision involve deep learning solutions which, given enough data, produce impressive performance in a variety of tasks, including those using biometric data [26–29]. But state-of-the-art solutions are usually based on deep Convolutional Neural Networks (CNN) with dozens of millions of parameters, and whose models typically have hundreds of megabytes, for example [30]. This makes unfeasible their applicability to mobile devices, both because of computational constraints, and of size limitations imposed by marketplaces to downloadable applications. If we look at state-of-the-art results with the database that we employ in the present paper [31–33] (Table 9), they all use very deep networks which would not be transferable to mobiles. Thus, models capable of operating under the restrictions of such environments are necessary. Another limitation is the lack of large databases of ocular images for soft-biometrics [15]. To overcome this, it is common to start with networks pre-trained on another tasks for which large databases exist. Examples include the generic ImageNet Challenge [34], as done for example in [19, 35], or face recognition datasets [32, 36]. Both are approaches that we follow in the present paper as well.

1.1 | Contributions

This article focuses on the use of smartphone ocular images to estimate age and gender. Partial faces can be expected in unconstrained environments, but also in controlled ones due to the use of masks, thus our focus on the ocular region as the only visible part of the face. To be clear, we have not employed images of people wearing masks, or occluded images, but we have cropped the ocular area from selfie face images. This also allows to compare the use of the entire face or only the ocular region with the same input data. A preliminary version appeared in a conference [4]. Here, we employ another database, Adience [21], consisting of Flickr images uploaded with smartphones that are jointly annotated with age and gender. It also has a more balanced distribution between classes. Given its in-the-wild nature, it provides a more demanding setup. In some other works (see Tables 1 and 2), images are taken in controlled environments, for example from face databases (such as MORPH or FERET), or using close-up capture typical of iris acquisitions (such as Cross-Eyed, GFI, UTIRIS, ND-Iris-0405, etc.).

Datasets for age and gender prediction from social media are still relatively limited [1]. To counteract over-fitting, some works use small CNNs of two or three convolutional layers trained from scratch [16, 17, 35, 37]. To be able to use more complex networks, one possibility is to pre-train them on a generic task for which large databases exist, like ImageNet [34]. This is done for example in [19, 35], and in the present paper. In the previous study, we employed CNNs pre-trained on ImageNet as well, and classification was done with Support Vector Machines (SVMs). In contrast, end-to-end training of the networks on the target domain is evaluated here too. Also, the present study evaluates networks pre-trained in a related task, face recognition [6, 52], where large databases are available. Since both tasks use the same type of input data, we aim at analysing if such face recognition pre-training can be beneficial for soft-biometrics. Other works have followed this strategy as well [32, 36], but they employ the entire face. Thus, to the best of our knowledge, taking advantage of networks pre-trained for face recognition for the task of ocular soft-biometrics can be considered novel.

Finally, this paper is oriented towards the use of smartphone images. This demands architectures capable of working in mobile devices, a constraint not considered in our previous study. The lighter CNNs that we employ [53, 54] have been proposed for common visual tasks in the context of the ImageNet challenge, and they have been bench-marked for face recognition as well [6, 55, 56]. To achieve less parameters

TABLE 1 Age prediction from ocular images. Only Adience contains selfie images captured with frontal smartphone cameras. See the text for details

Work	Year	Features	Database	Spectrum	Images	Eyes	Best Accuracy
[16]	2017	CNN	Adience	VIS	12,460	Both	46.97% ± 2.9 (exact), 80.96% ± 1.09 (1-off)
[37]	2014	23 sub-CNNs to face parts	MORPH	VIS	55,244	Face patches	MAE = 3.63 years
[17]	2019	4 sub-CNNs to face parts	Adience	VIS	19,370	Face patches	51.03% ± 4.63 (exact), 83.41% ± 3.17 (1-off).
[38]	2019	SURF/SVM-kNN	own	VIS	500	Both	96.57%
[4]	2020	CNN/SVM	LFW	VIS	12,007	One/Both	60.2/60% (exact)
This paper		CNN/SVM	Adience	VIS	11,299	One/Both	45.9/48.8% (exact), 83.1/86.2% (1-off)

Abbreviations: CNN, convolutional neural networks; MAE, Mean Absolute Error; SVM, Support vector machines.

TABLE 2 Gender prediction from ocular images. Only Adience and VISOB contain selfie images captured with frontal smartphone cameras. See the text for details

Work	Year	Features/Classifier	Database	Spectrum	Images	Eyes	Best Accuracy
[39]	2010	LBP/LDA, PCA, SVM	web data	VIS	936	Both	85%
[40]	2011	Shape features/MD, LDA, SVM	FRGC	VIS	800	One	97%
[41]	2012	ICA/NN	FERET	VIS	200	Both	90%
[42]	2016	HOG, LBP, LTP, WLD/SVM	group pictures	VIS	2921	Both	83%
[43]	2016	BSIF/SVM	BioCOP	NIR	3314	One	85%
[18]	2017	Textural descriptors/SVM, MLP	VISOB	VIS	1200	One	90.2%
[19]	2018	CNN/SVM, MLP, KNN, AdaBoost, CNN	VISOB	VIS	1200	One/Both	89.01/90.0
[44]	2017	Intensity, Texture, Shape/Random Forest	Cross-Eyed	VIS + NIR	3840	One	90%
[35]	2018	CNN/NN, CNN	GFI	NIR	4976	One	85.48%
[45]	2019	Intensity, Texture, Shape/SVM, ensembles	GFI, UTIRIS, Cross-Eyed, UNAB	VIS + NIR	11,973	One	89.22%
[46]	2019	SRCNN/Random Forest	CSIP, MICHE, MOBBIO, own	VIS	6450	One	90.15%
[47]	2018	CNN	GFI	NIR	3000	One/Both	85.06/87.26%
[48]	2017	Deep Class-Encoder	GFI, ND-Iris-0405	NIR	67,979		83.17%
[49]	2018	GIST perceptual descriptors	self-captured	multi-spectral	8320	One	81%
[50]	2019	BSIF, LBP, LPQ/SVM	BioCOP2009, Cosmetic Contact, GFI	NIR	51,006	One	86%
[20]	2019	compass LBP/SVM	Adience, cFERET, LFW, CUFS, CUFSF	VIS	1757	One/Both	84.06/83.27%
[51]	2019	ULBP + BSA/SVM	CASIA-Iris-Distance, MGBC	NIR, VIS	705	One/Both	66.67/78%
[4]	2020	CNN/SVM	LFW	VIS	12,007	One/Both	92.6/93.4%
This paper		CNN/SVM	Adience	VIS	11,299	One/Both	76.6/78.9%

Abbreviations: BSIF, Binarized Statistical Image Feature; CNN, Convolutional Neural Networks; HOB, histograms of oriented gradients; NIR, near-infrared; SVM, Support Vector Machines.

and faster processing while keeping accuracy, they use techniques such as point-wise convolution, depth-wise separable convolution, bottleneck layers, or residual connections [54]. The models obtained have a few megabytes (Table 3), in contrast to other popular models such as ResNet [30], which occupy dozens or hundreds of megabytes.

The contributions of this paper to the state-of-the-art are thus:

- We summarise related works in age and gender classification using ocular images.
- We apply two generic lightweight CNN architectures to the tasks of age and gender estimation. The networks, SqueezeNet [53] and MobileNetv2 [54], were proposed in the context of the ImageNet Challenge [34], where the networks are pre-trained with millions of images to classify thousands of generic object categories.

The networks proposed within ImageNet have been used in the literature as base models in many other recognition tasks [57], especially when available data is insufficient to train them from scratch.

TABLE 3 Networks evaluated in this paper. The vector size corresponds to the layer prior to the classification layer of each CNN (used for SVM training)

Network	Input Size	Conv Layers	Model Size	Para-metres	Vector Size
MobileNetv2 [54]	113 × 113	53	13 MB	3.5 M	1280
SqueezeNet [53]	113 × 113	18	4.41 MB	1.24 M	1000
MobileFaceNets [55]	113 × 113	50	4 MB	0.99 M	512
MobiFace [56]	113 × 113	45	11.3 MB	n/a	512
ResNet50 [52]	224 × 224	50	146 MB	25.6 M	2048
SENet50 [52]	224 × 224	50	155 MB	28.1 M	2048

Abbreviations: CNN, Convolutional Neural Networks; SVM, Support Vector Machines.

There is the assumption that architectures that perform well on a generic task like ImageNet will perform well on other vision tasks [58]. Thus, it is common to use ImageNet pre-trained networks just as fixed feature extractors, taking the output of the last layers as descriptor, and use it to train a classifier (like SVM) for the new task.

In some cases, the network is re-trained taking ImageNet weights as initialisation even if there is sufficient training data for the new task, since it can produce faster convergence than scratch initialisation [58].

The networks that we have selected for the present paper are two of the smallest generic architectures proposed within ImageNet, specifically tailored for mobile environments.

To be precise, the architectures employed were presented by their respective authors [53, 54] in the context of the ImageNet challenge, and here we apply them to the task of ocular soft-biometrics classification.

We have also implemented two existing lightweight architectures proposed specifically in previous studies for face recognition using mobile devices, MobileFaceNets [55] and MobiFace [56]. They are based on MobileNetv2, but with a smaller size and number of parameters.

- To assess if more complex networks can be beneficial, we also evaluate two CNN-based on the large ResNet50 model [30] and on Squeeze-and-Excitation (SE) blocks [59]. ResNet was also proposed within ImageNet, presenting the concept of residual connections to ease the training of CNNs. They have been also applied successfully to face recognition [52]. In this paper, we apply these existing architectures to soft-biometric classification.

Proposed without mobile restrictions in mind, they have significantly more parameters and size than the networks of the previous point (Table 3). However, as we have observed, it does not translate in superior performance, at least with the amount of training data available in this paper.

- The available networks are comprehensively evaluated for age and gender prediction with smartphone ocular images. For comparative purposes, we also use the entire face. To this aim, we use a challenging dataset, Adience [21], which consists of selfie images captured in real-world conditions with smartphones. To the best of our knowledge, this is the first work that compares the use of face and ocular images for age and gender prediction with this database. We also conduct experiments using two different ocular ROIs consisting of single eye images and combined descriptors from both eyes.
- Classification experiments with the networks are done in two ways: by using feature vectors from the layer prior to the classification layer, and then training a separate SVM classifier; and by training the networks end-to-end. Prior to this, the networks are initialised in different ways. First, we use the large-scale ImageNet pre-training [34], an approach followed in many other classification tasks [57]. It allows to use the network as feature extractor and simply train a classifier, or to facilitate end-to-end training if there is little data in the target domain [58]. Due to previous research [6, 52], the CNNs are also available after being fine-tuned for face recognition with two large databases [52, 60]. Even if face recognition is a different task, we hypothesise that such fine-tuning can be beneficial for soft-biometrics classification. Indeed, facial soft-biometrics indicators also allow to separate identities [9], so features learn for one task can aid the other. In addition, since the ocular region appears in face images, we speculate that networks trained for face recognition can benefit soft-biometric estimation using ocular images as well.
- Results of our experiments are reported in several ways. First, the accuracy of the networks is reported for the various initializations and classification options evaluated.

Convergence of the end-to-end training is also analysing by showing the training curves, including training and inference times.

Finally, t-SNE scatter plots of the vectors given by the last layer of the networks are also provided, showing progressive separation of the classes as the network progresses from a generic training (ImageNet) to an end-to-end training which also includes face recognition fine-tuning in the process.

The rest of the paper is organised as follows. A summary of related works in age and gender classification using ocular images is given in Section 2. Section 3 then describes the networks employed. The experimental framework, including database and protocol, is given in Section 4. Extensive experimental results are provided in Section 5, followed by conclusions in Section 6.

2 | RELATED WORKS ON AGE AND GENDER CLASSIFICATION USING OCULAR IMAGES

Pioneering studies of age or gender estimation from RGB ocular smartphone images were carried out by Rattani et al. [16, 18, 19]. Previously, near-infrared (NIR) iris images for age

estimation were employed, taking advantage of available iris databases [61, 62]. These studies used geometric or textural information, attaining an accuracy of ~64%. Gender estimation from iris texture had been also proposed [40, 63–69], reaching an accuracy over 91% [66]. These early works followed the pipeline of iris recognition systems, so soft-biometrics classification was done extracting features from the segmented iris region (even if the surrounding ocular region is visible). Later works, mentioned below, have incorporated the ocular region to the analysis, even if the images are captured using traditional NIR iris sensors. Before the availability of specific ocular databases, it was also common to crop the ocular region from face databases like FRGC [40], FERET [41], web-retrieved data [39], or pictures of groups of people [42]. There are also works using the entire face [70, 71] but due to space, we concentrate only on ocular images. Tables 1 and 2 summarise previous work on age and gender prediction. Only two databases (Adience and VISOB) are captured with frontal smartphone cameras (selfie-like). Databases like MORPH, LFW, FRGC, FERET, etc. contain face images, of which the ocular region is cropped. Other databases are of ocular images captured with digital cameras (Cross-Eyed), or iris images with NIR sensors (e.g. BioCOP, GFI, UTIRIS, UNAB, ND-Iris-0405).

Age classification from smartphone ocular images is carried out in [16] using their own proposed CNN. To avoid overfitting, they use a small sequential network with 3 convolutional and 2 fully-connected layers (41,416 learnable parameters), which takes as input a crop of 32 × 92 pixels of the two eyes. Experiments are done with 12,460 images of the Adience benchmark [21], which is also employed in the present paper. The database contains face images, so the ocular ROI is extracted by landmark localisation with the DLib detector [72]. To simulate selfie-like case, only frontal images are retained. The reported accuracy is 46.97 ± 2.9 (exact) and 80.96 ± 1.09 (1-off).

A set of works apply a patch-based approach for age estimation, in which crops of face regions are used [17, 37]. In [37], the authors use 23 patches around facial landmarks to fed 23 small CNNs (of 3 convolutional layers), each CNN specialised in one patch. Landmarks are detected using Active Shape Models. The patches operate at different scales, with the larger scale covering the entire face, and their outputs are connected together in a fully-connected layer. Therefore, the algorithm rely on combining regions of the entire face. Experiments are done with 55,244 images of the MORPH database, which includes age labels from 16 to 77 years. The Mean Absolute Error (MAE) is of 3.63 years. The authors also found that patches capturing smaller areas of the face give better results than patches that capture big areas, although the best accuracy is obtained when all scales are combined. Inspired by [37], the authors of [17] use a CNN architecture of 4 branches, having 4.8 M learnable parameters. Each branch, of just 3 convolutional layers, is specialised on one patch around the eyebrows, eyes, nose or mouth. These regions are detected using the OpenFace and DLib detectors [72]. The branches are then connected to a fully-connected layer. During training, the loss of each branch and the loss of entire network are summed up. However, each branch estimator is not used at inference time, but only the concatenated soft-max, so the system relies on the availability of all regions. The approach is evaluated with 19,370 in-plane aligned images of Adience. The accuracy is 51.03 ± 4.63 (exact) and 83.41 ± 3.17 (1-off). The authors also removed different branches to evaluate its contribution, noticing that the absence of eyes and mouth contributed most to reducing the accuracy (specially the eyes). This supports studies like the present one, which concentrates on the ocular region as the most prominent facial part for soft-biometrics.

In a recent work [38], the authors use SURF (Speeded Up Robust Features) to detect key-points and extract features from the ocular region. Then, a hybrid SVM-kNN classifier is applied. With a small database of 500 images, they achieve an age accuracy of 96.57%.

More recently, we applied CNNs pre-trained on Imagenet to the tasks of age, gender and ethnicity [4] with 12,007 images of the Labelled Faces in the Wild (LFW) database. One of the CNNs is also pre-trained for face recognition, as in the present work, although in [4] it did not prove to be an advantage. We extract features of different regions (face, eyes and mouth) using intermediate layers of the networks identified in previous works as providing good performance in ocular recognition [73, 74]. Then, we train SVMs for classification. In overall terms, the accuracy using ocular images only drops ~2%–4% in comparison to the entire face. The reported accuracy is 95.8%/64.5% in gender/age estimation (entire face), 92.6%/60.2% (ocular images), and 90.5%/59.6% (mouth images). The approach is also evaluated against two commercial off-the-shelf systems (COTS) that employ the whole face, which are outperformed in several tasks.

Regarding gender estimation, the work [43] pioneered the use of different regions around the iris for prediction. It uses Binarised Statistical Image Feature (BSIF) texture operator, and SVM as classifier. Data consists of 3314 NIR images of the BioCOP database. The work found that the entire ocular region provides the best accuracy (~85%) and excluding the iris has a small impact (~84%). On the other hand, using only the iris texture pushes down accuracy to less than 75%, highlighting the importance of the periocular region. The first study making use of selfie ocular images was presented in [18]. It evaluates several textural descriptors in combination with SVMs and Multi-layer Perceptrons (MLPs). They use 1,200 selfie images of the VISOB database captured with 3 smartphones. The left and right eyes are cropped to 240 × 160 pixels with the Viola-Jones eye detector. The work reports results for each smartphone, with the best accuracy being 90.2%. Later, the same authors evaluated pre-trained and custom CNNs on the same database [19]. The very deep VGG and ResNet networks (pre-trained on ImageNet), along with a custom CNN of 3 convolutional layers, are employed. Experiments are conducted on single eye images (of 120 × 123) and on strips of both eyes (120 × 290). The pre-trained networks are used to extract feature vectors (from the last layer before soft-max) that feed an external classifier. The authors evaluated SVMs, MLPs, K-nearest neighbours (KNN), and Adaboost. The best

accuracy (90.0 ± 1.35) was obtained with pre-trained networks and both eyes. The custom CNN is just behind (89.60 ± 2.91). Using only one eye, the best accuracy is 89.01 ± 1.30 (pre-trained CNNs) and 87.41 ± 3.07 (custom CNN).

In [44], Tapia and Viedma address gender classification with RGB and NIR ocular images. They employ pixel intensity, texture features (Uniform Local Binary Patterns, ULBP), and shape features (Histograms of Oriented Gradients, HOG) at different scales. Classification is done with Random Forest using 3840 images of the Cross-Eyed database. Among the different findings, we can highlight that: it is better to extract features at different scales than in a single scale only, and the fusion of features from RGB and NIR images improves accuracy. They also compare the extraction of features from the iris texture or the surrounding ocular area, finding that the ocular area is best, attaining an accuracy of 90%.

In subsequent works, Viedma et al. [35, 45] study gender classification with NIR ocular images. In [35], they train two small CNNs of 2 and 3 convolutional layers from scratch. They also use the very deep VGG-16, VGG-19 and Resnet-50 architectures (pre-trained on ImageNet). As in [19], the pre-trained networks are used as fixed feature extractors to feed a classifier (a dense neural network in this case). The authors also fine-tune these pre-trained networks by freezing the initial convolutional layers. Experiments are done with 4976 images of 120 × 160 from the GFI database, which are augmented using several spatial transformations. The custom CNNs were found to perform better (best accuracy 85.48%). They also observed (via activation maps of the networks) that the ocular area that surrounds the iris is the most relevant to classify gender, more than the iris area itself. In [45], the authors employ the same features as in [44], together with SVMs and nine ensemble classifiers. They use 4 databases with gender information: GFI (4976 images), UTIRIS (389), Cross-Eyed (3840) and UNAB-gender (2768). The best accuracy is 89.22%, achieved by selecting features from the most relevant regions using XgBoost. As in [35], the relevant features are spread throughout the whole ocular area with the exception of the iris.

Later on, authors from the same group [46] applied super-resolution convolutional networks (SRCNNs) to counteract scale variability in the acquisition of selfie ocular images in real conditions. They use 4 databases of VIS images: CSIP (2004 images), MOBBIO (800), MICHE (3196) and a self-captured one (450). Classification is done with Random Forest. The work shows that increasing resolution (2× and 3×) improves accuracy, achieving 90.15% (right eye) and 87.15% (left eye). In another paper [47], they applied a small CNN of 4 convolutional layers, both trained separately for each eye, and for the fused left-right eye images. They use 3000 NIR images of the GFI database, showing that training the network separately for each eye is best (87.26% accuracy).

The work [48] applies a variant of an auto-encoder (Deep Class-Encoder) to predict gender and race using NIR iris images of 48 × 64 pixels. The databases employed for gender experiments are GFI (2999 images) and ND-Iris-0405 (64,980 images). The best gender accuracy is 83.17% (GFI) and 82.53% (ND-Iris-0405).

In [49], they use GIST perceptual descriptors with weighted kernel representation to carry out gender classification from images captured in 8 different spectral bands simultaneously. To this aim, the authors use a spectral imaging camera. With a self-captured database of 104 ocular instances (10 different captures per instance, totalling 104 × 10 × 8 images), they achieve an average accuracy of 81%.

In [50], the authors use NIR ocular images to estimate gender and race. They apply typical iris texture descriptors used for recognition (Binarised Statistical Image Feature, BSIF, Local Binary Patterns, LBP, and Local Phase Quantization, LPQ) with SVM classifiers. Three datasets are used: Bio-COP2009 (41,830 images), Cosmetic Contact (4,200), and GFI (4,976). The gender accuracy from a single eye image is of 86%. The study also confirms previous research that showed that excluding the iris region provides greater accuracy.

The authors of [20] apply a patch-based approach for gender estimation with 10 crops around landmarks (left eye, right eye, complete eye region, lower nose, lip, left face, right face, forehead and upper nose). Then, compass LBP features are extracted from each region, and classified with one SVM per region. Finally, the classification scores of all regions are combined with a genetic algorithm. Experiments are done with Adience (1757 images), colour FERET (987), LFW (5749) and two sketch datasets, CUFS (606) and CUFSF (987 sketches from colour FERET). The best accuracy is 95.75% (colour FERET). The performance on Adience using the whole face is 87.71%. The authors also study each facial region individually on the Adience database, with an accuracy of 84.06% (one eye) and 83.27% (both eyes). Other regions of the face provide lower accuracy (73.95%–82.71%), with the lip region providing 78.25%. This supports the findings of our previous study, which revealed the eye region as having superior accuracy than other regions of the face [4].

Lastly, in [51], it is proposed a multimodal system that fuses features from the face and ocular regions. They use 300 NIR images of CASIA-Iris-Distance, and 405 VIS images of the MBGC database (one third are faces, one third are left eye, and one third are right eye images). As features, they employ ULBP (with overlapping blocks), and Backtracking Search Algorithm (BSA) to reduce feature dimensionality. Classification is done via SVM by combining the features of each available face region. With CASIA-Iris-Distance, the accuracy of the individual regions is 82.00 ± 0.82 (face), 66.00 ± 0.75 (left eye), 62.00 ± 0.70 (right eye), and 78.00 ± 0.91 (both eyes). After the fusion, accuracy goes up to 88.00 ± 0.94. With MGBC, the accuracy is 81.67 ± 1.03 (face), 66.67 ± 0.65 (left eye), 66.67 ± 0.73 (right eye), 74.17 ± 0.35 (both eyes), and 92.51 ± 0.68 (fusion).

3 | CNNS FOR SOFT-BIOMETRICS PREDICTION

We extract features from the face, left and right ocular regions (Figure 1, top) using different CNN architectures (Table 3). Two light-weight pre-trained generic architectures, SqueezeNet

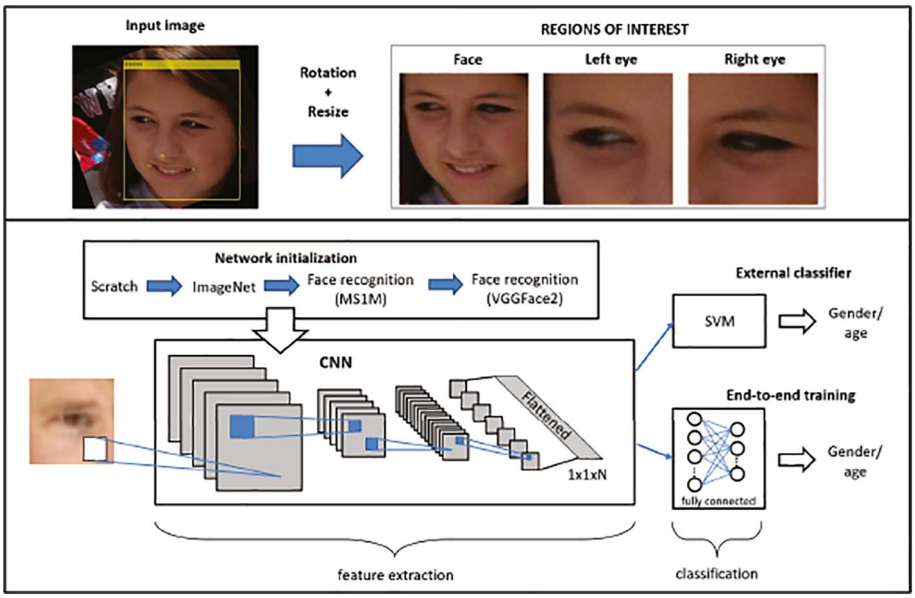


FIGURE 1 Top: Extraction of the regions of interest. Bottom: Soft-biometrics classification framework

and MobileNetv2, are used for feature extraction and classification.

- *SqueezeNet* [53] is one of the early networks designed to reduce the number of parameters and model size. The authors proposed the use of squeeze and expand modules that follow the bottleneck concept. First, dimensionality is reduced with 1 × 1 point-wise convolutions (squeeze or bottleneck layer), followed by a layer with a larger amount of filters (expansion layer), which includes 3 × 3 filters too. The network uses late down-sampling, since keeping large activation maps should lead to a higher accuracy. With only 1.24 M parameters, 4.6 MB and 18 convolutional layers, it matched AlexNet accuracy on ImageNet with 50x fewer parameters.
- *MobileNetv2* [54] employs depth-wise separable convolutions and inverted residual structures to achieve a light architecture. Depth-wise separable convolution works in two stages, first performing filtering with a single filter per input channel, followed by a 1 × 1 point-wise convolution that linearly combine the channels. In the case of 3 × 3 filters, this reduces computations by a factor of eight or nine compared to a standard full convolution, with a small cost in accuracy [75]. Inverted residual structures, also called bottleneck residual blocks with expansion, consists of first expanding the number of channels with 1 × 1 point-wise filters. Then, they are processed with a large amount of 3 × 3 depth-wise separable filters. Finally, the number of channels is reduced again with 1 × 1 point-wise filters. A shortcut (residual) connection is added between the input and the output of such structure to improve the ability of a gradient to propagate across layers. This network has 3.5 M parameters, a size of 13Mb and 53 convolutional layers.

The original SqueezeNet and MobileNetv2 are modified to employ an input size of 113 × 113 × 3. The stride of the first convolutional layer is changed from 2 to 1, so the networks can remain unchanged (more importantly, we can use ImageNet weights). We have also implemented two lightweight architectures proposed specifically for face recognition using mobile devices. They are MobileFaceNets [55] and MobiFace [56]. Both are based on MobileNetV2, but with smaller expansion factors on bottleneck layers to make the network smaller. They also employ a reduced input image size of 113 × 113 × 3. *MobileFaceNets* has 0.99 M parameters, 50 convolutional layers, and 4 MB. It uses Global Depth-wise Convolution (GDC) to substitute the standard Global Average Pooling (GAP) at the end of the network. The motivation is that GAP treats all pixels of the last channels equally, but in face recognition, the centre and corner pixels should be weighted differently It also uses PReLU as non-linearity, and fast down-sampling at the beginning. *MobiFace* [56] also employs fast down-sampling and PReLU, but the authors changed GAP by a fully-connected layer to allow learning of different weights for each spatial region of the last channels. This network has a size of 11.3 MB and 45 convolutional layers.

Finally, we evaluate the large models of [52] for face recognition. They use ResNet50 [30] and SE-ResNet50 (abbreviated as SENet50) [59] as backbone architectures, with an input size of 224 × 224 × 3. ResNet networks presented the idea of residual connections to ease CNN training. Followed later by many (including MobileNetV2), residual connections allow much deeper networks. The network employed here, ResNet50, has 50 convolutional layers, but there are deeper ResNets of even 1001 layers [76]. The Squeeze-and-Excitation (SE) blocks [59], on the other hand, explicitly model channel relationships to adaptively recalibrate channel-wise feature responses. SE blocks can be integrated with other architectures, such as ResNet, to improve its representation power.

4 | EXPERIMENTAL FRAMEWORK

4.1 | Database

We use the Adience benchmark [21], designed for age and gender classification. The dataset consists of Flickr images uploaded automatically with smartphones. Some examples are shown in Figure 2. Given the uncontrolled nature of such images, they have high variability in pose, lightning, etc. The downloaded dataset includes 26,580 images from 2,284 subjects. To simulate selfie captures, we removed images without frontal pose, resulting in 11,299 images. They are then rotated w.r.t. the axis crossing the eyes, and resized to an inter-eye distance of 105 pixels (average of the database). Facial landmarks are extracted using the MTCNN detector [77]. Then, a face image of 224 × 224 is extracted around the mass centre of the landmarks, together with the ocular regions (of 113 × 113 each). The breakdown of images into the different classes is given in Table 4.

4.2 | Protocol

The Adience benchmark specifies a 5-fold cross-validation protocol, with splits pre-selected to avoid images from the same Flickr album appearing in both training and testing sets in the same fold. Given a test fold, classification models are trained with the remaining four folds. Classification results, therefore, consist of mean accuracy and standard error over the five folds. Following [21], we also provide the 1-off age classification rate, in which errors of one age group are considered correct classifications. The training folds are augmented by mirroring the images horizontally. In addition, the illumination of each image is varied via gamma correction with $\gamma = 0.5$, 1, 1.5 ($\gamma = 1$ logically leaving the image unchanged). This way, from a single face or ocular image, we generate 6 training images, with which we expect to counteract over-fitting and accommodate variations in illumination. Finally, when feeding the CNNs, images are resized to the corresponding input size indicated in Table 3.

Classification is done in two ways (Figure 1, bottom): *i*) by training a linear SVM [78] using feature vectors extracted from the CNNs, and *ii*) by training the CNNs end-to-end. Prior to training, the CNNs are initialised in different ways, as will be explained in Section 5. To train the SVMs, we use vectors from the layer prior to the classification layer, with the size of the feature vectors given in Table 3. When there are more than two classes (age classification), a one-vs-one multi-class approach is used. For every feature and N classes, $N(N-1)/2$ binary SVMs are used. Classification is based on which class has most number of binary classifications towards it (voting scheme). Regarding end-to-end training, we change the last fully connected layer of each network to match the number of classes (2 for gender, 8 for age). Batch-normalisation and dropout at 50% is added before the fully-connected layer to counteract over-fitting. The networks are trained using soft-max as loss function and Adam as optimiser, with mini-batches of 128 (we also tried SGDM initially, but Adam provided better accuracy overall, therefore we skipped SGDM). The learning rate is 0.001. During training, 20% of images of each class are set aside for validation in order to detect over-fitting and stop training accordingly. When the networks are initialised from scratch, training is stopped after five epochs. In all other cases, training is stopped after two epochs. Experiments have been done in stationary computers running Ubuntu, with an i9-9900 processor, 64 Gb RAM, and two NVIDIA RTX 2080 Ti GPUs. We carry out training using Matlab r2020b, while the implementations of ResNet50 and SENet50 are run using MatConvNet.

5 | RESULTS

5.1 | Pre-trained CNN models

The SqueezeNet, MobileNetv2 and ResNet50 CNNs are available pre-trained on the large-scale ImageNet dataset [34].

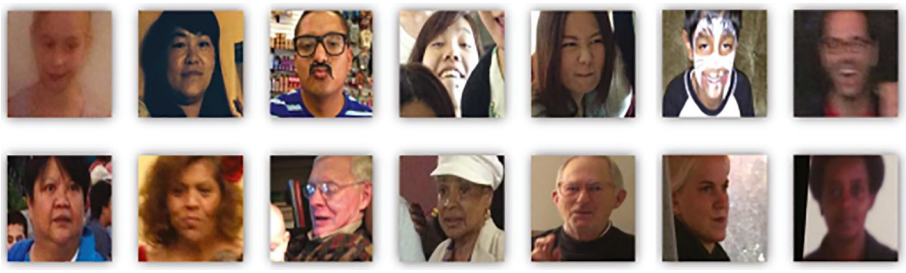


FIGURE 2 Images from the Adience database (from [21])

TABLE 4 Breakdown of face images of the database into the different classes

Male	Female	0–2	4–6	8–13	15–20	25–32	38–43	48–53	60–99
5,353	5,946	1,003	1,546	1,665	1,095	3,298	1,578	551	563

They are also available after they have been fine-tuned for face recognition using two large face databases [6, 52]. To do so, the networks are trained for biometric identification on the MS-Celeb-1M database [60] (MS1M for short), and then fine-tuned on the VGGFace2 database [52]. The images of these databases, downloaded from the Internet, show large variations in pose, age, ethnicity, lightning and background (see Figure 3). MS1M has 10M images from 100k celebrities (with an average of 81 images per subject), while VGGFace2 has 3.31 M images of 9131 subjects (362.6 images per subject). Fine-tuned ResNet50 and SENet50 models are made available by the authors of [52], initialised from scratch. SqueezeNet and MobileNetv2 are trained by us as described in [6], initialised using ImageNet weights, and producing the models trained with MS1M, and then on VGGFace2. MobileFaceNets and MobiFace are also trained by us with the same protocol, but initialised from scratch.

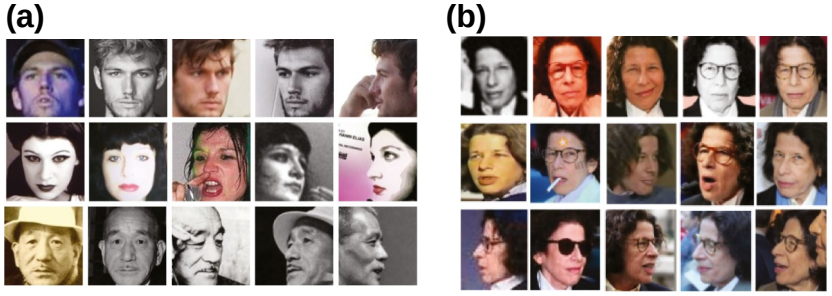
Table 5 shows the classification performance obtained with these pre-trained networks, and using SVM as classifier, according to the protocol of Section 4.2. For each CNN, different possibilities based on the available pre-training are reported. We provide age and gender classification results using as input either the whole face or the ocular region. The columns named 'ocular' refer to the left and right eyes separately (each image is classified independently), while 'ocular L + R' refer to the combination of both eyes (by averaging the CNN descriptors before calling the SVM classifier).

In the majority of networks, a better accuracy is obtained after the CNNs are fine-tuned for face recognition on MS1M or VGGFace2. Also, it is better in general if the networks have undergone the double fine-tuning, first on MS1M, and then on VGGFace2. This goes in line with the experimentation of [6, 52], which showed that face recognition performance could be improved after this double fine-tuning. These results also show that a CNN trained for face recognition can be beneficial for soft-biometrics classification too, even if just the ocular region is employed. Given that facial soft-biometric cues can be used for identity recognition as well [9], features learnt for recognition are expected to carry soft-biometrics information, and vice-versa. The only exception is ResNet50, where a better accuracy in general is obtained only with the ImageNet training. This shows as well that even ImageNet training can be beneficial for soft-biometrics (as shown in our previous paper too [4]), since the accuracy of ResNet50 on ImageNet is similar or better in some cases than the accuracy obtained with other CNNs after they are fine-tuned for face recognition. With SENet50 we cannot draw any special conclusion since there is only one pre-training available. What it can be said is that it performs worse than ResNet50, even if in face recognition tasks, SENet50 is better (as reported in [6, 52]).

Regarding face versus ocular classification, there is no clear winner when the networks are only trained with ImageNet. Gender accuracy is marginally better with the entire face, with the biggest difference observed with ResNet50 (78.3% vs. 71.9%). Regarding age, the ocular area shows comparable accuracy, and even better in some cases, for example: 38.7% versus 40.4% (ResNet50, exact accuracy), or 36.6% versus 37.8% (MobileNetv2, exact accuracy). The indicated ocular accuracy refers to both eyes ('ocular L + R'), which is observed to improve by 3%-4% in comparison to using one eye only. This comparable accuracy between face and ocular regions is a very interesting result. Since the networks are trained for a generic recognition task like ImageNet, and not particularly optimised to the use of facial or ocular images, we can safely assume that the ocular region is a powerful region for soft-biometrics estimation, and comparable to the entire face. This is in line with our previous findings as well [4].

When the networks are fine-tuned for face recognition with MS1M or VGGFace2, accuracy with the entire face becomes substantially better (sometimes by ∼15%). Still, accuracy with the ocular area is improved as well. This may be because it appears in the training data, although in a small portion of the image. This leads us to think that accuracy with the ocular area could be made comparable if the CNNs are fine-tuned for ocular recognition instead.

Lastly, from the results of Table 5, we cannot conclude that one CNN is better than other. A good CNN for gender is MobileNetv2, which is the best with the ocular region, and its face accuracy is good as well. For age classification, MobiFace stands out. It should be highlighted though that the difference



MS1M images of three users (by row) and three viewpoints (by column: frontal (1-2), three-quarter (3-4), and profile (5)).

VGGFace2 images from three viewpoints (frontal, three-quarter, and profile, arranged by row). Image from [38].

FIGURE 3 Example images of the MS1M and VGGface2 databases

TABLE 5 Accuracy of gender and age estimation using pre-trained CNN models and SVM classifiers. The best results with each network are marked in bold. For each column, the best accuracy is highlighted with a grey background

Pre-training				Gender			Age					
							Face		Ocular		Ocular L + R	
ImageNet	MS1M	VGGFace2	Network	Face	Ocular	Ocular L + R	Exact	1-Off	Exact	1-Off	Exact	1-Off
X			ResNet50	78.3 ± 1.6	**69.2 ± 1.7**	**71.9 ± 1.7**	38.7 ± 5.6	78.2 ± 2.5	**37.1 ± 4.5**	**73.8 ± 3.5**	**40.4 ± 5.6**	**77.4 ± 3.8**
	X	X	ResNet50	**82 ± 2.2**	66.6 ± 1.7	69.9 ± 1.7	**53.8 ± 4.8**	**93.4 ± 1.1**	33.2 ± 4.1	70.2 ± 2.9	37.3 ± 5.2	75.1 ± 4
	X	X	SENet50	**81.5 ± 3.6**	**64.2 ± 0.6**	**67.2 ± 1**	**52.4 ± 4.7**	**92.7 ± 2.3**	**33.2 ± 3.9**	**68.6 ± 2.9**	**37.9 ± 4.8**	**73.6 ± 3.6**
X			MobileNetv2	72.3 ± 1.8	65.6 ± 2.1	68.7 ± 2.7	36.6 ± 3.7	73.8 ± 2.6	34.3 ± 4.1	69.3 ± 4	37.8 ± 4.9	73.2 ± 3.8
X	X		MobileNetv2	81.4 ± 1.2	**72.2 ± 2.5**	**75.1 ± 2.7**	49.6 ± 5.3	88.8 ± 2.9	**39.5 ± 3.9**	**76.1 ± 2.8**	**43.7 ± 5.1**	**80.3 ± 3.4**
X	X	X	MobileNetv2	**82.3 ± 2.4**	69.3 ± 2.1	72.1 ± 2.5	**49.7 ± 5**	**90.4 ± 1.9**	36.6 ± 4.1	74.7 ± 3.1	40.9 ± 4.5	79.1 ± 3.7
X			SqueezeNet	74.2 ± 2.6	67.1 ± 1.4	69.7 ± 2.2	39.3 ± 4.6	77.7 ± 2.5	35 ± 3.8	70.3 ± 3.6	37.7 ± 5.1	73.9 ± 4
X	X		SqueezeNet	78.8 ± 2.6	70.2 ± 2.3	73.2 ± 2.5	46.8 ± 3.3	85.3 ± 2.6	38 ± 3.3	73.5 ± 3	42.2 ± 4.2	77.4 ± 3.4
X	X	X	SqueezeNet	**82.9 ± 2**	**71.2 ± 2**	**73.8 ± 1.6**	**48 ± 5.8**	**88.2 ± 2.2**	**38.3 ± 4.1**	**74.2 ± 3.1**	**42.4 ± 4.8**	**77.9 ± 3.4**
	X		MobileFaceNets	80.9 ± 1.3	70.4 ± 3	73 ± 3.3	50.7 ± 4.4	88.4 ± 2.2	37.5 ± 3.7	72.6 ± 2.8	41.4 ± 5.5	76.7 ± 3.5
	X	X	MobileFaceNets	**84.3 ± 0.8**	**70.9 ± 2.2**	**74 ± 2.8**	**53.7 ± 3.9**	**91.1 ± 2.6**	**38.5 ± 5.1**	**74.7 ± 2.9**	**42.2 ± 6.2**	**78.1 ± 3.6**
	X		MobiFace	79.3 ± 2.2	71.1 ± 1.4	73.5 ± 1.2	49 ± 5.3	87.3 ± 2.6	40.6 ± 4	76.2 ± 3	44.6 ± 4.8	79.6 ± 3.4
	X	X	MobiFace	**81.9 ± 1.2**	**71.5 ± 2.4**	**74.4 ± 2.2**	**51.6 ± 4.8**	**90.2 ± 2.1**	**41.2 ± 5.1**	**77 ± 3.2**	**45.1 ± 6.3**	**80.4 ± 3.7**

Abbreviations: CNN, Convolutional Neural Networks; SVM, Support Vector Machines.

between CNNs is 2%–3% or less in most columns. This is interesting, considering that the networks differ in size, sometimes substantially (Table 3). It is especially relevant to observe that ResNet50 and SENet50 do not outperform the others, even if the input image size and the network complexity is higher. A final observation is that gender classification is more accurate in general than (exact) age classification. Being a binary classification, gender may be regarded as less difficult than age recognition, which has eight classes. In addition, we have employed the same database for both tasks, so age classes contain less images for training. If we consider the 1-off age rate, on the other hand, age accuracy becomes better than gender accuracy.

5.2 | CNN models fine-tuned for soft-biometrics classification

Four networks are further fine-tuned to do the classification end-to-end, according to the protocol of Section 4.2. We keep only the small CNNs, since they will be less prone to over-fitting, given the reduced number of images in comparison to, for example, face recognition [6, 52]. Table 6 shows the classification results considering different pre-training, including from scratch.

As in the previous section, a better accuracy is obtained with the CNNs that are fine-tuned first for face recognition on MS1M or VGGFace2, rather than only on ImageNet. However, in this case, it is sufficient if the networks are just fine-tuned on MS1M. Training from scratch produces the worst results, suggesting that the amount of training data is not yet sufficient in comparison to other domains. A way to overcome such problem is to train the networks first in other tasks for which large-scale databases are available, as we do in this paper. A generic task like ImageNet can be useful [57], producing better results than in the network is just trained from scratch. But according to our experiments, a better solution is to use a task for which similar training data is employed, such as face recognition.

In Table 7 and Figure 4, we provide the training curves over two epochs, and training/inference times of the different models (pre-trained on MS1M, which is the model that provides the best accuracy overall in Table 6). Due to space constraints, we show only the results over the first fold of the database. Figure 4 shows that most models converge over the first epoch (first half of the horizontal axes), with the validation accuracy showing little improvement over the second epoch. The horizontal axes of the periocular plots reach a higher value because for each face image, there are two separate ocular images, so the number of iterations is doubled. It can be also seen that the validation accuracy after the second epoch (red and blue for gender and age, respectively) is similar in most cases to the accuracy reported in Table 6, that is 70%–80% for gender estimation, and 40%–50% for age estimation. Regarding training times, ocular obviously takes double due to the duplication of images. Also, gender and age training takes comparatively the same time for each CNN, given that the same images are used, but divided into different classes. The depth of each network (convolutional layers, see Table 3) correlates with the training time. The lightest network (SqueezeNet) takes the least time, while the deepest ones (MobileNetv2 and MobileFaceNets) take the longest. Inference times have been computed with the CPU to simulate lack of graphical power. Still, times are in the order of milliseconds, showing correlation with the depth of the CNN as well.

Regarding face versus ocular classification, the same conclusions than in the previous section apply. When the networks have not seen such type of images before (scratch or ImageNet pre-training), face and ocular images produce comparable performance. The difference is just 2%–3% with most networks, the only exception being SqueezeNet, for which face is better than ocular by up to 10%. On the other hand, when the CNNs are fine-tuned for face recognition, then accuracy with the entire face becomes substantially better, although accuracy with the ocular area is improved as well.

In contrast to the previous section, Table 6 shows MobileNetv2 as the clear winner, producing the best accuracy in all tasks. This network is the more complex of all four (see Table 3), which may explain its superiority. The other networks should not be dismissed though, since their accuracy is just 2%–3% below in most cases, so a lighter network provides just a slightly worse accuracy.

Comparing the results of Tables 5 and 6, we observe that the best accuracy per network (bold elements) is in general equal or better in the experiment of this section (Table 6), except the 1-off age accuracy. Even if all networks improve the exact age estimation to a certain extent, accuracy in this task is still below 50%, which may be a sign that more training data would be desirable. The degradation in 1-off age accuracy may be another sign of over-fitting. The network that benefits the most from the training of this section is MobiletNetv2, with improvements of 3%–5%. MobileFaceNets and MobiFace show improvements of 2%–3% in the majority of tasks. Squeezenets shows marginal improvements in gender estimation, with some improvements of 2%–3% in age estimation only.

To further evaluate the improvements given by the end-to-end training of this section, we have removed the fully-connected layers of each network, and trained SVMs instead for classification, as in Section 5.1. Results are shown in Table 8. Interestingly, gender accuracy is degraded, but age shows some improvement in exact estimation (1%–3%), and a substantial improvement in 1-off estimation (18%–20%). For example, the best 1-off age face/ocular accuracy is 91.3%/86.2%, surpassing the best gender results obtained in this paper. The results of Table 8 suggest that SVM is a better classifier in the difficult age estimation task. Table 8 also shows the superiority of MobileNetv2, having the best accuracy in nearly all tasks.

Finally, we evaluate the benefits of the progressive fine-tuning proposed by showing in Figure 5 the scatter plots created by t-SNE [79] of the vectors provided by each network just before the classification layer. The t-SNE settings are exaggeration = 4, perplexity = 30, learning rate = 500. For MobileNetv2, we show results after different network training

TABLE 6 Accuracy of gender and age estimation using CNN models trained end-to-end. The best results with each network are marked in bold. For each column, the best accuracy is highlighted with a grey background. Underlined elements indicate that its accuracy is worse than the corresponding combination in Table 5

Pre-training				Gender			Age					
							Face		Ocular		Ocular L + R	
ImageNet	MS1M	VGGFace2	Network	Face	Ocular	Ocular L + R	Exact	1-Off	Exact	1-Off	Exact	1-Off
from scratch			MobileNetv2	72.3 ± 1.2	67.5 ± 3.1	70 ± 3	38.8 ± 4	61.5 ± 2.6	36.4 ± 5.1	59.2 ± 3.7	39 ± 5.3	61.5 ± 3.5
X			MobileNetv2	79.6 ± 2.6	74 ± 1.7	76.6 ± 2.1	41.9 ± 3.2	65.7 ± 3.5	39.8 ± 5.1	62 ± 4.1	42.3 ± 5.5	64.3 ± 4.1
X	X		MobileNetv2	**85.3 ± 5.4**	**76.6 ± 3.3**	**78.9 ± 3.7**	**52 ± 5.7**	**73.9 ± 3.8**	**45.9 ± 4.3**	**66.9 ± 3.4**	**48.4 ± 4.3**	**69.2 ± 3.1**
X	X	X	MobileNetv2	80.1 ± 4.1	74 ± 3	76.9 ± 2.9	48.6 ± 4.3	71.7 ± 3.6	41 ± 4.2	63.3 ± 3.1	43.4 ± 4.5	65.6 ± 3.3
from scratch			SqueezeNet	62.8 ± 9.4	52.8 ± 2.3	52.8 ± 2.3	39.6 ± 4.5	61.2 ± 3	34.7 ± 5.5	55.7 ± 3.7	35.9 ± 6	57.1 ± 4.4
X			SqueezeNet	69.6 ± 9.4	62.8 ± 3.5	63.2 ± 3.9	44.6 ± 6.2	65.3 ± 5.1	34.2 ± 7.8	57.5 ± 2	35.2 ± 8.7	58.6 ± 2.8
X	X		SqueezeNet	**82.1 ± 2.1**	70.8 ± 5.5	72.6 ± 6.9	**51.1 ± 4.9**	**72.7 ± 3.3**	**42.9 ± 4.2**	62.9 ± 4.4	**45.4 ± 4.6**	64.7 ± 4.3
X	X	X	SqueezeNet	79.4 ± 3.6	**71.6 ± 1.6**	**73.5 ± 1.3**	48.3 ± 4.6	70.6 ± 3.2	40.8 ± 6.5	**63.3 ± 5.4**	43.2 ± 7.2	**65.4 ± 6.1**
from scratch			MobileFaceNets	74.1 ± 5.7	69.5 ± 2.6	72.3 ± 3	40.8 ± 5.5	63.4 ± 3.5	34.6 ± 4.4	57.2 ± 4.2	37.2 ± 4.9	59.5 ± 4.4
	X		MobileFaceNets	**84.1 ± 2.7**	**74.3 ± 2.9**	**76.9 ± 2.6**	**50.7 ± 3.6**	**73.1 ± 1.9**	**43 ± 4.1**	64.9 ± 3.5	**45.5 ± 4.8**	67 ± 3.6
	X	X	MobileFaceNets	82 ± 3.4	73.8 ± 2.7	76 ± 3.1	49 ± 5.7	71.2 ± 2.6	42.2 ± 4.3	**65.5 ± 5.3**	44.8 ± 4.7	**67.9 ± 5.5**
from scratch			MobiFace	69.1 ± 4.3	64.1 ± 3.2	65.8 ± 4.4	35.2 ± 5.5	58 ± 4	28.1 ± 3.7	51.9 ± 4.4	29.2 ± 4.6	52.8 ± 4.8
	X		MobiFace	**82.9 ± 2.2**	72.9 ± 2.2	75.3 ± 2.4	**50.6 ± 5.2**	**72.2 ± 4.7**	**41 ± 5.4**	**62 ± 3.7**	**43.7 ± 5.8**	**64.7 ± 3.9**
	X	X	MobiFace	81.3 ± 2.2	**73 ± 2**	**75.8 ± 2.6**	44.3 ± 4.2	69.2 ± 1.4	37.5 ± 4.5	59.7 ± 4.1	40.2 ± 5.1	62 ± 4.6

Abbreviations: CNN, Convolutional Neural Networks.

TABLE 7 Training and inference times of the networks evaluated in this paper. Training times correspond to the plots shown in Figure 4. The computers used equip a Intel i9-9990 CPU @ 3.1 GHz, 64 Gb RAM and two NVIDIA RTX 2080 Ti GPUs. Inference times are computed in CPU mode

	Training end to end (mm:ss)				
Network	Gender Face	Gender Ocular	Age Face	Age Ocular	Inference
MobileNetv2 [54]	54:02	100:22	53:50	90:25	19.7 ms
SqueezeNet [53]	37:27	71:49	37:25	72:39	6.2 ms
MobileFaceNets [55]	68:53	122:35	62:27	114:58	29.2 ms
MobiFace [56]	42:31	81:17	36:25	75:13	17.5 ms

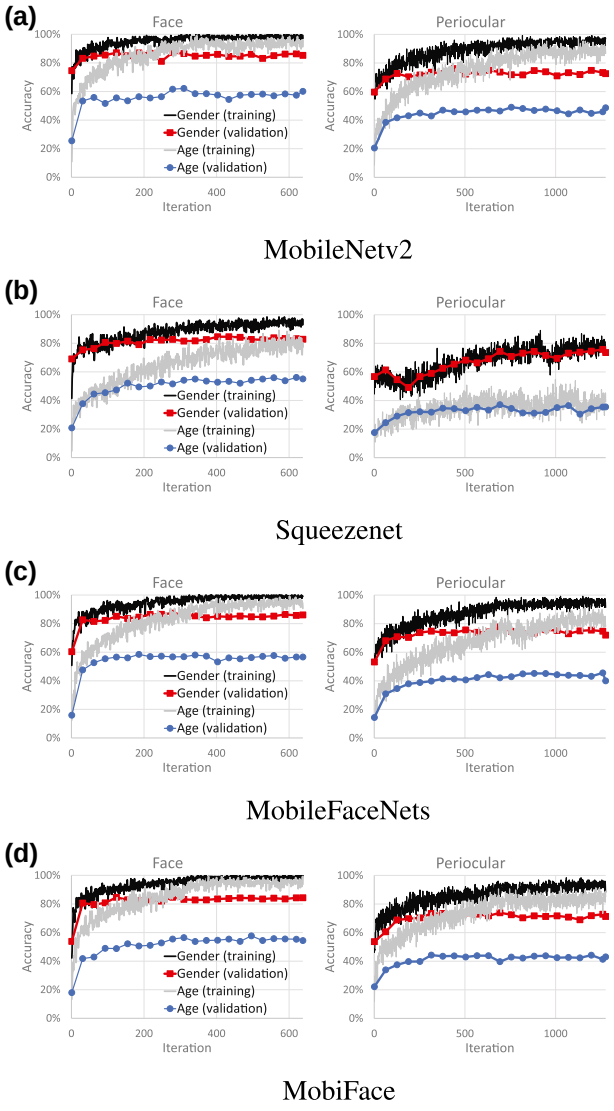


FIGURE 4 Training progress of the convolutional neural networks (CNNs) for soft-biometrics (with networks pre-trained on MS1M for face recognition, the case that provides the best accuracy in Table 6). All plots correspond to 2 training epochs over the training set of the first fold of Adience

(left column, top to bottom): *i*) ImageNet (generic pre-training), *ii*) ImageNet + MS1M (fine-tuning to face recognition), and *iii*) ImageNet + MS1M + Adience (fine-tuning to soft-biometrics classification). It can be seen that as the plots progresses from top to bottom, the cluster of each class tend to separate more from the others. For gender classification, the blue and red dots form two distinct clouds in the third row. For age classification, clusters of age 0–2 (red), 4–6 (orange) and 8–13 (green) appear nearby, and in progressive (circular) trajectory towards the light blue clusters (15–20 and 25-32 age groups). Then, the dark blue clusters (38–43 and 48–53) appear, and finally, the 60–99 group (magenta). This progressive ordering of age groups in the feature space reflects the expected correlation in human age [1], with adjacent age groups being closer to each other, and non-adjacent age groups appearing more distant. Similar class separations after ImageNet + MS1M + Adience training is also observed with the other three networks (right column). On the contrary, in training *i* and *ii* with MobileNetv2, the clusters are spread across a common region, without a clear separation among them, especially with only ImageNet training. Male/female clusters are intertwined, forming a circle-like cloud, and the same happens with age groups. Light blue (young adults), dark blue (middle age) and magenta (old age) dots are spread across the same region. Even red and orange dots (children) sometimes appear in opposite extremes of the circle-like shape.

5.3 | Summary and comparison with previous works

Table 9 shows a summary of the best reported accuracy of the two previous sub-sections (cells highlighted with a grey background in Tables 5–8). For reference, the performance of other works using the same database for ocular age or gender estimation is also shown [16, 20]. Most of the literature making use of the Adience database employ full-face images, with the best published accuracy shown also at the bottom of Table 9. To identify these works [31–33], we have reviewed all citations to the papers describing the database [21, 80] reported by IEEEXplore (circa 305 citations), and selected the ones with the best published accuracy for each column. It must be noted that although the Adience database is divided in pre-defined folds, the works of Table 9 may not necessarily employ the same amount of images per fold, so results are not completely comparable.

In gender estimation, we do not outperform the related work that uses the ocular region [20]. It should be highlighted that the latter uses 1757 images (see Table 2), while we employ 11,299. We outperform previous age accuracy using the ocular region [16], which uses a set of comparable size (12,460 images). To prevent over-fitting, the paper [16] uses a small custom CNN trained from scratch with images of 32 × 92 (crop of the two eyes). In contrast, our networks are pre-trained on several tasks, including generic object classification [34], and face recognition [6, 52], which seems a better option. Our input image size is also bigger (113 × 113).

TABLE 8 Accuracy of gender and age estimation using CNN models fine-tuned for soft-biometrics classification and SVM classifiers. The best results with each network are marked in bold. For each column, the best accuracy is highlighted with a grey background

Pre-training				Gender			Age					
							face		Ocular		Ocular L + R	
ImageNet	MS1M	VGGFace2	Network	face	Ocular	Ocular L + R	Exact	1-Off	Exact	1-Off	Exact	1-Off
X	X		MobileNetv2	**78.5 ± 2.2**	**70.4 ± 1.9**	**73.3 ± 2.8**	**53.4 ± 5.3**	**91.3 ± 2.2**	**44.8 ± 4.5**	**82.9 ± 3.1**	**48.6 ± 5**	**86.2 ± 3.1**
X	X	X	MobileNetv2	75.6 ± 2.2	67.6 ± 2	70.6 ± 2.7	52 ± 5.8	89.7 ± 2.5	43.5 ± 5.1	82.2 ± 3.4	47.6 ± 5.4	85.3 ± 3.3
X	X		SqueezeNet	76.2 ± 1.9	**67.6 ± 2.7**	**69.9 ± 2.9**	**51.3 ± 4.8**	**89.3 ± 2.8**	**43.5 ± 4.9**	79.8 ± 3.3	47 ± 5.2	82.7 ± 3.9
X	X	X	SqueezeNet	**77.1 ± 1.7**	66.8 ± 2.1	69 ± 2.3	50.6 ± 5.7	**89.3 ± 2.5**	43.3 ± 3.5	**80.9 ± 2.9**	**47.4 ± 4.7**	**83.9 ± 3.4**
	X		MobileFaceNets	76.7 ± 1.1	69.1 ± 2	71.8 ± 2.1	**52.7 ± 3.2**	89.8 ± 2.3	43.3 ± 3.8	81.4 ± 3.4	47.1 ± 4.3	84.6 ± 3.7
	X	X	MobileFaceNets	**78.4 ± 2.6**	**69.9 ± 2.2**	**72.8 ± 2.7**	52.4 ± 5	**91.2 ± 2.5**	**45.1 ± 4.3**	**83.1 ± 3.7**	**48.8 ± 5.1**	**85.4 ± 4.1**
	X		MobiFace	78.1 ± 1.8	**69.1 ± 1.7**	**71.7 ± 2.4**	49.5 ± 5.4	88.2 ± 2.5	41.3 ± 3.4	79.4 ± 2.8	45.6 ± 4.2	83.1 ± 3.3
	X	X	MobiFace	**78.3 ± 1.2**	68.6 ± 1.8	**71.7 ± 2.3**	**51.5 ± 6**	**90.2 ± 2.7**	**43.2 ± 5.2**	**81.3 ± 3.7**	**47 ± 6.2**	**84.7 ± 3.8**

Abbreviations: CNN, Convolutional Neural Networks; SVM, Support Vector Machines.

Compared to works using the full face, we do not outperform them either [31–33]. In gender estimation, we obtain an accuracy ∼8% behind the best method [31]. The latter uses the very deep VGG19 CNN (535 MB, 144 M parameters), which is much more complex than the networks employed here (Table 3). The size of the input image is 448 × 448, which is much bigger than ours. Also, a saliency detection network is trained first on the PASCAL VOC 2012 dataset to detect the regions of interest ('person' or 'face' pixels) and indicate the classification CNNs the pixels to look at. In age estimation, our accuracy is more competitive, ∼4% behind the best method in 1-off classification [33], although the exact accuracy is still way behind the best result [32]. The work [33] combines residual networks (ResNet) or residual network of residual networks (RoR) with Long Short-Term Memory (LSTM) units. First, a ResNet or a RoR model pre-trained on ImageNet is fine-tuned on the large IMDB-WIKI-101 dataset (500k images with exact age label within 101 categories) for age estimation. Then, the model is fine-tuned on the target age dataset to extract global features of face images. Next, to extract local features of age-sensitive regions, a LSTM unit is presented to find such age-sensitive region. Finally, age group classification is conducted by combining the global and local features. The size of the input image is 224 × 224, and the best reported accuracy is obtained with a ResNet152 network as base model (214 MB), an even deeper network that the ResNet50 evaluated in the present paper. The work [32] follows an approach similar to ours to prevent over-fitting. They use the very deep VGG-Face CNN (516 MB) [81], which is trained to recognise faces using ∼1 million images from the Labelled Faces in the Wild and YouTube Faces datasets. To fine-tune the model for age classification, the CNN is frozen, and only the fully connected layers are optimised. The network uses images of 224 × 224 for training. For testing, they use images of 256 × 256, of which 5 images of 224 × 224 are extracted (four corners and centre). Then, the five images are fed into the CNN, and the softmax output vectors are averaged. This combination method is also followed by the authors of Adience [80], showing some improvement in comparison to the centre crop only.

We lastly report the detail of gender and age estimation of each class for our approach (Tables 10 and 11). We also include (when available) the details of other approaches of Table 9. It can be observed that gender recognition is relatively equal between classes (1%–2% of variation around the overall accuracy), which can be a result of the classes being well balanced in the database (Table 5). Regarding age, the accuracy between classes is more variable. It may may be a product of the classes being less balanced, although there are not always correlation between class representation and accuracy. It can also be seen that all methods show the same relative performance among classes. This includes other works [32, 80], even if they are based on different networks or training strategies, suggesting that some classes may be more difficult. The classes with the worst accuracy are 38–43 and 48–53 in the majority of columns, but the class 48–53 is much less represented in the database. The class 15–20 also has comparatively low

TABLE 9 Summary of the best reported accuracy of the experiments of this paper. The table also includes results of recent works using the same database. Different papers may not employ exactly the same amount of images per fold, so results are not completely comparable. The best results of our experiments are marked in bold. For each column, the best accuracy is highlighted with a grey background

				Age					
				Face		Ocular		Ocular L + R	
Method	Gender face	Ocular	Ocular L + R	Exact	1-Off	Exact	1-Off	Exact	1-Off
Best of Table 5	84.3 ± 0.8	72.2 ± 2.5	75.1 ± 2.7	**53.8 ± 4.8**	**93.4 ± 1.1**	41.2 ± 5.1	77 ± 3.2	45.1 ± 6.3	80.4 ± 3.7
Best of Table 6	**85.3 ± 5.4**	**76.6 ± 3.3**	**78.9 ± 3.7**	52 ± 5.7	73.9 ± 3.8	**45.9 ± 4.3**	66.9 ± 3.4	48.4 ± 4.3	69.2 ± 3.1
Best of Table 8	78.5 ± 2.2	70.4 ± 1.9	73.3 ± 2.8	53.4 ± 5.3	91.3 ± 2.2	45.1 ± 4.3	**83.1 ± 3.7**	**48.8 ± 5.1**	**86.2 ± 3.1**
Best of [16] (2017)	-	-	-	-	-	-	-	46.97 ± 2.9	80.96 ± 1.09
Best of [20] (2019)	87.71%	84.06	83.27	-	-	-	-	-	-
Best of [21] (2014)	77.8 ± 1.3	-	-	45.1 ± 2.6	79.5 ±0 .4	-	-	-	-
Best of [80] (2015)	86.8 ± 1.4	-	-	50.7 ± 5.1	84.7 ± 2.2	-	-	-	-
Best of [31] (2019)	93.52	-	-	-	-	-	-	-	-
Best of [32] (2020)	-	-	-	70.96	92.7	-	-	-	-
Best of [33] (2020)	-	-	-	67.83 ± 2.98	97.53 ± 0.59	-	-	-	-

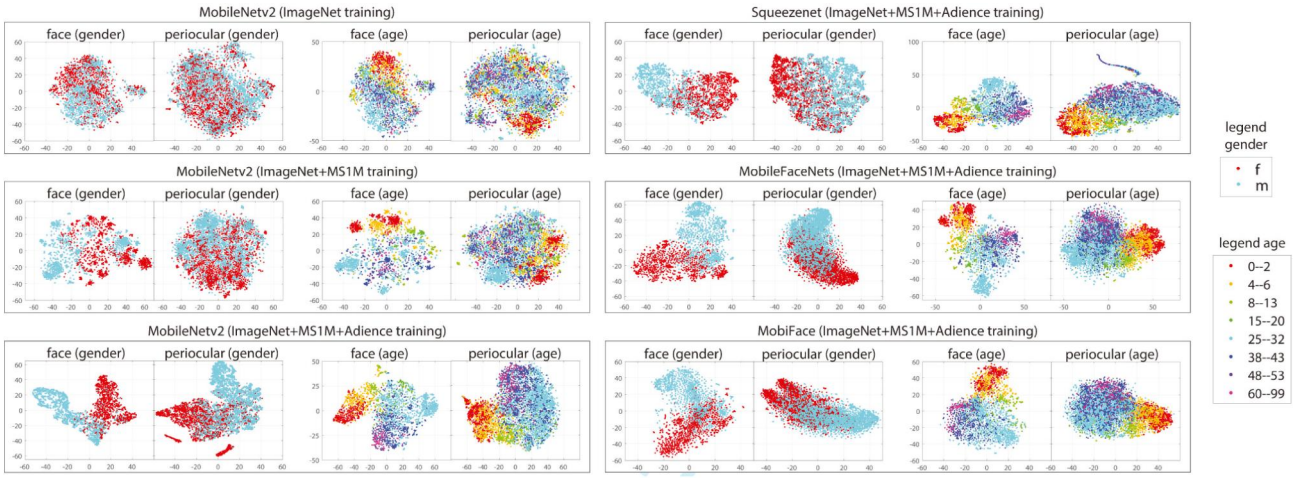


FIGURE 5 Scatter plots by t-SNE of the vectors from the layer prior to the classification layer of each convolutional neural network (CNN). The vector size (dimensionality) of each CNN is shown in Table 3. All plots are generated with the test set of the first fold of the Adience database. Best in colour and with zoom

performance. On the other hand, other classes with low representation (0–2 and 60–99) have better performance, and in some cases, 0–2 even shows the best accuracy. The most represented class (25–32) does correlate with the best accuracy in some cases, and its performance is among the best in most columns.

6 | CONCLUSION

We are interested in lightweight network architectures capable of providing age and gender recognition using selfie ocular images. The literature review suggests that many of the proposed methods use data captured in controlled ways, either cropped from RGB face databases or from iris databases that employ close-up near-infrared sensors. Also, to be able to operate in mobile devices, the models have to be sufficiently small, making infeasible the use of very large Convolutional Neural Networks (CNNs) that provide state-of-the-art results in related tasks such as identity or expression recognition [27, 29]. Their typical size (hundreds of megabytes) prevent their incorporation in downloadable mobile applications, where the entire file typically cannot exceed 100 Mb. Accordingly, we have adapted very light models of a few megabytes [53–56] to operate with small ocular images. The networks employed can also provide inference in <30 ms on a CPU, so a mobile device with sufficient power should be able to run them in real-time too. To counteract over-fitting due to

TABLE 10 Detail of gender estimation results (columns 2–4 refer to the cases with best overall accuracy in our experiments)

Class	face	Ocular	Ocular L + R	face [20]
Overall	85.3	76.6	78.9	87.71
Female	86.1	78.1	80.4	86.80
Male	84.1	74.7	77.1	88.69

TABLE 11 Detail of age estimation results (columns 2–4 refer to the cases with best overall accuracy in our experiments). For each column, the best accuracy is highlighted with a grey background, and the worst accuracy is marked in bold

Class	face	Ocular	Ocular L + R	face [80]	face [32]
Overall	53.8	45.9	48.8	50.7	70.96
0–2	76.2	57	47	69.9	98.9
4–6	63.2	24	64.4	57.3	79.7
8–13	52.3	60.5	49	55.2	75.2
15–20	36.2	29.7	23.4	23.9	68.1
25–32	64.1	62	68.6	61.3	47.3
38–43	43.6	**19.4**	35.2	29.3	67.5
48–53	**30.4**	32.7	**16.8**	**14.6**	**41.7**
60–99	49.3	39.8	27	35.7	79.8

the lack of very large selfie datasets for age and gender prediction, we use architectures pre-trained on the ImageNet Challenge [34], where the networks have learnt to classify thousands of generic object categories by using millions of training images. We also exploit the availability of very large face recognition databases [52, 60]. Due to previous research [6, 52], the networks are fine-tuned first for face recognition. We hyphothesize that such large-scale fine-tuning can be beneficial for soft-biometrics classification too, since both tasks uses the same type of input data.

Experiments are done with 11,299 images of the Adience benchmark, which contains in-the-wild smartphone images uploaded to Flickr. The networks are evaluated for age and gender prediction using images of the ocular region. For comparison, they are also evaluated with the entire face. Classification is done in two ways: by extracting feature vectors from the layer prior to the classification layer of the network, and then training a SVM classifier; and by training the network end-to-end. We also compare different network initialisation, including from scratch, with ImageNet weights, and fine-tuned for face recognition (as mentioned above).

In our experiments, training from scratch provides the worst results, suggesting that training data is not yet sufficient compared to other domains. Initialising the networks with a generic task for which large databases exists (like ImageNet) is more efficient [57], as done by in another soft-biometrics works too [19, 35]. But in most cases, the best accuracy is obtained when the CNNs are fine-tuned first for face recognition. This is also observed in the t-SNE plots of the vectors given by the networks, where the classes appear more separated after such face recognition pre-training. Such phenomenon is observed even if only the ocular region is used for soft-biometrics estimation, which we attribute to the ocular region appearing in face images, so it is 'seen' by the networks previously. Identity and soft-biometrics are inter-related tasks, since they use the same input data. Indeed, soft-biometrics can aid identity recognition as well [9], so it is expected that one task benefits the other. Regarding face versus ocular classification, there is no clear winner when the networks are initialised with ImageNet, as observed in previous research too [4]. In such a case, the networks are trained for a generic task, without a particular optimization to facial or ocular images. Thus, we can consider the ocular region as a powerful standalone region for soft-biometrics, comparable to the entire face. On the contrary, when the networks are initialised with face recognition weights, soft-biometrics classification with the entire face becomes substantially better (although accuracy with the ocular area is improved as well). Our interpretation is that since the ocular region appears in portions of the face image, such initialisation also benefits the ocular soft-biometric task, although to a lesser extent. We believe that if the networks are fine-tuned for ocular recognition instead, ocular soft-biometric classification would become comparable to the entire face, as observed with the agnostic ImageNet initialisation.

Regarding absolute numbers, our best accuracy is 85.3%/93.4% in gender/1-off age estimation with the entire face, and 78.9%/86.2% with the combination of the two eyes. In gender ocular recognition, we do not outperform the best accuracy of the literature with the Adience database [20], although the mentioned work uses 10% of the images that we employ in this paper. In age ocular recognition, we outperform previous research [16]. The majority of research with this database is done with full-face images, but existing papers producing state-of-the-art results [31–33] (Table 9) all use very deep networks, which would not be transferable to mobile devices.

As future work, we are looking into fine-tuning the networks for ocular recognition, given that such area can be cropped from face databases. This way, we expect to increase ocular soft-biometrics accuracy by transfer-learning, as observed after the networks are trained for face recognition. Also, this work has simultaneously addressed age and gender recognition with a single database, but larger repositories of unconstrained data containing only one of these indicators are becoming available, for example [28, 82]. This would allow to separately address each task with bigger datasets, although it would hinder another direction that we want to pursue, which is joint-estimation of both indicators. We foresee that improvements can be obtained by sharing weights between the networks, since a single facial feature can carry information not only about identity, but about different soft-biometrics at the same time. One plausible direction to overcome this would be to train the networks on larger databases for each task, as done by works that focus on gender [31] or age estimation [33] separately, and then combine them together onto a database

labelled with several soft-biometric indicators simultaneously. Freezing initial layers after the networks have been pre-trained in a related task (such as face recognition) can be another approach to counteract the lack of sufficient data in the target database, as done by other studies as well [32]. Age estimation using ocular data also deserves extra attention. The exact accuracy is still low in comparison to gender estimation. With the employed database, state-of-the-art accuracy is 93.52% (gender) versus 70.96% (age), see Table 9. In ocular works with another databases (Tables 1 and 2), a gender accuracy of 90%–95% is common, while exact age estimation barely reaches 60%. We expect to achieve improvements in this direction with larger facial repositories [83].

ACKNOWLEDGMENTS

Part of this research has been enabled by a visiting position of F. Alonso-Fernandez at the University of the Balearic Islands (UIB), funded by the UIB visiting lecturers programme. Authors F. Alonso-Fernandez, K. Hernandez-Diaz and J. Bigun would like to thank the Swedish Research Council for funding their research. Authors F. J. Perales and S. Ramis would like to thank the projects PERGAMEX RTI2018-096,986-B-C31 (MINECO/AEI/ERDF, EU) and PID2019-104829RA-I00/AEI/10.13039/501100011033 (MICINN).

ORCID

Fernando Alonso-Fernandez https://orcid.org/0000-0002-1400-346X

REFERENCES

1. Sun, Y., et al.: Demographic analysis from biometric data: achievements, challenges, and new frontiers. IEEE Trans. Pattern Anal. Mach. Intell. 40(2), 332–351 (2018)
2. Dantcheva, A., Elia, P., Ross, A.: What else does your biometric data reveal? a survey on soft biometrics. IEEE Trans. Inform. Forensic Secur. 11(3), 441–467 (2016)
3. Alonso-Fernandez, F., Bigun, J.: A survey on periocular biometrics research. Pattern Recogn Lett. 82, 92–105 (2016)
4. Alonso-Fernandez, F., et al.: Soft-biometrics estimation in the era of facial masks. In: International Conference of the Biometrics Special Interest Group (BIOSIG), pp. 1–6 (2020)
5. Ramachandra, R., Busch, C.: Presentation attack detection methods for face recognition systems: A comprehensive survey. vol. 50, edn. 1. Association for Computing Machinery, New York (2017)
6. Alonso.Fernandez, F., et al.: Lihtweight face verification across different poses for mobile platforms. In: Proc. IAPR TC4 Workshop on Mobile and Wearable Biometrics, WMWB, in conjunction with International Conference on Pattern Recognition (ICPR). IEEE, (2020)
7. Akhtar, Z., et al.: Biometrics: in search of identity and Security (Q & A). IEEE MultiMedia. 25(3), 22–35 (2018)
8. Rattani, A., et al. (eds.) Introduction to selfie biometrics, pp. 1–18. Springer International Publishing, Cham (2019)
9. Gonzalez-Sosa, E., et al.: Facial soft biometrics for recognition in the wild: recent works, annotation, and COTS evaluation. IEEE Trans. Inform.Forensic Secur. 13(8), 2001–2014 (2018)
10. Dantcheva, A., et al.: Bag of soft biometrics for person identification. Multimed Tools Appl. 51, 739–777 (2011)
11. Jain, A.K., Ross, A., Prabhakar, S.: An introduction to biometric recognition. IEEE Trans. Circuits Syst. Video Technol. 14(1), 4–20 (2004)
12. Tome, P., et al.: Soft biometrics and their application in person recognition at a distance. IEEE Trans. Inform. Forensic Secur. 9(3), 464–475 (2014)
13. Dantcheva, A., et al.: Search pruning in video surveillance systems: efficiency-reliability tradeoff. In: IEEE International Conference on Computer Vision Workshops (ICCV Workshops), pp. 1356–1363. IEEE, (2011)
14. Macedo, J., et al.: A benchmark methodology for child pornography detection. In: 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), pp. 455–462. IEEE, (2018)
15. Rattani, A., et al. (eds.). In:Soft-biometric attributes from selfie images, pp. 213–225. Springer International Publishing, Cham (2019)
16. Rattani, A., Reddy, N., Derakhshani, R.: Convolutional neural network for age classification from smart-phone based ocular images. In: IEEE International joint conference on biometrics (IJCB), pp. 756–761. IEEE, (2017)
17. de Assis Angeloni, M., de Freitas Pereira, R., Pedrini, H.: Age estimation from facial parts using compact multi-stream convolutional neural networks. In: IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), pp. 3039–3045. IEEE, (2019)
18. Rattani, A., Reddy, N., Derakhshani, R.: Gender prediction from mobile ocular images: a feasibility study. In: IEEE International Symposium on Technologies for Homeland Security (HST), pp. 1–6. IEEE, (2017)
19. Rattani, A., Reddy, N., Derakhshani, R.: Convolutional neural networks for gender prediction from smartphone-based ocular images. IET Biom. 7(5), 423–430 (2018)
20. Bhattacharyya, A., et al.: Recognising gender from human facial regions using genetic algorithm. Soft Computing. 23 (2019)
21. Eidinger, E., Enbar, R., Hassner, T.: Age and gender estimation of unfiltered faces. IEEE Trans. Inform. Forensic Secur. 9(12), 2170–2179 (2014).
22. Samangouei, P., Patel, V.M., Chellappa, R.: Attribute-based continuous user authentication on mobile devices. In: IEEE 7th International Conference on Biometrics Theory, Applications and Systems (BTAS), pp. 1–8. IEEE, (2015)
23. Antal, M., Nemes, G.: Gender recognition from mobile biometric data. In: IEEE 11th International Symposium on Applied Computational Intelligence and Informatics (SACI), pp. 243–248. IEEE, (2016)
24. Jain, A., Kanhangad, V.: Investigating gender recognition in smartphones using accelerometer and gyroscope sensor readings. In: International Conference on Computational Techniques in Information and Communication Technologies (ICCTICT), pp. 597–602. IEEE, (2016)
25. Alonso-Fernandez, F., et al.: Super-resolution for selfie biometrics: introduction and application to face and Iris. In: Rattani, A., Derakhshani, R., Ross, A. (eds.) Super-resolution for selfie biometrics: Introduction and application to face and iris, pp. 105–128. Springer International Publishing, Cham (2019)
26. Sundararajan, K., Woodard, D.L.: Deep learning for biometrics: a survey. ACM Comput. Surv. 51(3), 1–34 (2018)
27. Guo, G., Zhang, N.: A survey on deep learning based face recognition. Comput. Vis. Image Understand. 189, 102805 (2019)
28. Carletti, V., et al.: Age from faces in the deep learning revolution. IEEE Trans. Pattern. Anal. Mach. Intell. 42(9), 2113–2132 (2020)
29. Li, S., Deng, W.: Deep facial expression recognition: a survey, IEEE Trans. Affective Comput. 1 (2020)
30. He, K., et al.: Deep residual learning for image recognition. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778. IEEE, (2016)
31. Fang, J., et al.: Muti-stage learning for gender and age prediction. Neurocomputing. 334, 114–124 (2019)
32. Gyawali, D., et al.: Age range estimation using mtcnn and vgg-face model. In: 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), pp. 1–6. IEEE, (2020)
33. Zhang, K., et al.: Fine-grained age estimation in the wild with attention lstm networks. IEEE Trans. Circuits Syst. Video Technol. 30(9), 3140–3152 (2020)

34. Russakovsky, O., et al.: Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. 115(3), 211–252 (2015)
35. Viedma, I., Tapia, J.: Deep gender classification and visualization of near-infra-red periocular-iris images. In: IEEE International Conference on Image Processing, Applications and Systems (IPAS), pp. 73–78. IEEE, (2018)
36. Ozbulak, G., Aytar, Y., Ekenel, H.K.: How transferable are cnn-based features for age and gender classification? In: 2016 International Conference of the Biometrics Special Interest Group (BIOSIG), pp. 1–6. IEEE, (2016)
37. Yi, D., Lei, Z., Li, S.Z.: Age estimation by multi-scale convolutional network. In: Cremers, D. et al. (eds.) Asian Conference on Computer Vision, ACCV, pp. 144–158. Springer International Publishing, Cham (2015)
38. Kamarajugadda, K.K., Polipalli, T.R.: Extract features from periocular region to identify the age using machine learning algorithms. J. Med. Syst. 43(196) (2019)
39. Merkow, J., Jou, B., Savvides, M.: An exploration of gender identification using only the periocular region. In: Proceedings of International Conference on Biometrics: Theory Applications and Systems (BTAS), pp. 1–5. IEEE, (2010)
40. Dong, Y., Woodard, D.L.: Eyebrow shape-based features for biometric recognition and gender classification: a feasibility study. In: International Joint Conference on Biometrics (IJCB), pp. 1–8. IEEE, (2011)
41. Kumari, S., Bakshi, S., Majhi, B.: Periocular gender classification using global ICA features for poor quality images. In: International Conference on Modelling Optimization and Computing, ICMOC, vol. 38, pp. 945—951. Elsevier B.V. (2012)
42. Castrillón, S.M., et al.: On using periocular biometric for gender classification in the wild. Pattern Recogn Lett. 82, 181–189 (2016). an insight on eye biometrics
43. Bobeldyk, D., Ross, A.: Iris or periocular? exploring sex prediction from near infrared ocular images. In: International Conference of the Biometrics Special Interest Group (BIOSIG), pp. 1–7. IEEE, (2016)
44. Tapia, J., Viedma, I.: Gender classification from multispectral periocular images. In: IEEE International Joint Conference on Biometrics (IJCB), pp. 805–812. IEEE, (2017)
45. Viedma, I., et al.: Relevant features for gender classification in nir periocular images. IET Biom. 8(5), 340–350 (2019)
46. Tapia, J., et al. (eds.) In: Sex-classification from cellphones periocular iris images, pp. 227–242. Springer International Publishing (2019)
47. Tapia, J., Aravena, C.C.: Gender classification from periocular nir images using fusion of cnns models. In: IEEE 4th International Conference on Identity, Security, and Behavior Analysis (ISBA), pp. 1–6 (2018)
48. Singh, M., et al.: Gender and ethnicity classification of iris images using deep class-encoder. In: IEEE International joint conference on biometrics (IJCB), pp. 666–673. IEEE, (2017)
49. Raja, K. B., Raghavendra, R., Busch, C.: Fused spectral features in kernel weighted collaborative representation for gender classification using ocular images. In: 3rd International Conference on Computer Vision and Image Processing, pp. 131–143. Springer Singapore, Singapore (2020)
50. Bobeldyk, D., Ross, A.: Analyzing covariate influence on gender and race prediction from near-infrared ocular images. IEEE Access. 7, 7905–7919 (2019)
51. Eskandari, M., Sharifi, O.: Effect of face and ocular multimodal biometric systems on gender classification. IET Biom. 8(4), 243–248 (2019)
52. Cao, Q., et al.: A dataset for recognising faces across pose and age'. In: 13th IEEE International Conference on Automatic Face and Gesture Recognition (FG). pp. 67–74. IEEE, (2018)
53. Iandola, F.N., et al.: Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <1mb model size. CoRR (2016). abs/1602.07360. http://arxiv.org/abs/1602.07360
54. Sandler, M., et al.: Inverted residuals and linear bottlenecks. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4510–4520. IEEE, (2018)
55. Chen, S., et al.: Efficient cnns for accurate real-time face verification on mobile devices. CoRR (2018). abs/1804.07573. http://arxiv.org/abs/1804.07573
56. Duong, C.N., et al.: Mobiface: A lightweight deep learning face recognition on mobile devices. In: Proceedings of IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS) (2019)
57. Razavian, A.S., et al.: Cnn features off-the-shelf: an astounding baseline for recognition. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPRW, pp. 512–519. IEEE, (2014)
58. Kornblith, S., Shlens, J., Le, Q.V.: Do better imagenet models transfer better? In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2656–2666. IEEE, (2019)
59. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, (2018)
60. Guo, Y., et al.: Ms-celeb-1m: a dataset and benchmark for large-scale face recognition. In: Leibe, B., et al. (eds.) Computer Vision – ECCV 2016, pp. 87–102. Springer International Publishing, Cham (2016)
61. Erbilek, M., Fairhurst, M., Abreu, M.C.D.C.: Age prediction from iris biometrics. In: 5th International Conference on Imaging for Crime Detection and Prevention (ICDP), pp. 1–5. IET, (2013)
62. Sgroi, A., Bowyer, K.W., Flynn, P.J.: The prediction of old and young subjects from iris texture. In: International Conference on Biometrics (ICB), pp. 1–5. IEEE, (2013)
63. Thomas, V., et al.: Learning to predict gender from iris images. In: First IEEE International Conference on Biometrics: Theory, Applications, and Systems (BTAS), pp. 1–5. IEEE, (2007)
64. Lagree, S., Bowyer, K.W.: Predicting ethnicity and gender from iris texture. In: IEEE International Conference on Technologies for Homeland Security (HST), pp. 440–445 (2011)
65. Da Costa-Abreu, M., Fairhurst, M., Erbilek, M.: Exploring gender prediction from iris biometrics. In: International Conference of the Biometrics Special Interest Group (BIOSIG), pp. 1–11. IEEE, (2015)
66. Tapia, J.E., Perez, C.A., Bowyer, K.W.: Gender classification from iris images using fusion of uniform local binary patterns. In: Agapito, L., Bronstein, M.M., Rother, C. (eds.) Computer Vision - ECCV 2014 Workshops, pp. 751–763. Springer International Publishing (2015)
67. Tapia, J.E., Perez, C.A., Bowyer, K.W.: Gender classification from the same iris code used for recognition. IEEE Trans. Inform. Forensic Secur. 11(8), 1760–1770 (2016)
68. Kuehlkamp, A., Becker, B., Bowyer, K.: Gender-from-iris or gender-from-mascara? In: IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1151–1159. IEEE, (2017)
69. Tapia, J.: Gender classification from near infrared iris images. In: Rathgeb, C., Busch, C. (eds.) Iris and periocular biometric recognition. Institution of Engineering and Technology (2017)
70. Angulu, R., Tapamo, J.R., Adewumi, A.O.: Age estimation via face images: a survey. J. Image. Video. Proc. 42 (2018)
71. Osman, O.F., Yap, M.H.: Computational intelligence in automatic face age estimation: a survey. IEEE Trans. Emerg. Top. Comput. Intell. 3(3), 271–285 (2019)
72. King, D.E.: Dlib-ml: a machine learning toolkit. J. Mach. Learn. Res. 10, 1755–1758 (2009)
73. Hernandez, D.K., et al.: Periocular recognition using CNN features off-the-shelf. In: Proceedings of International Conference on Biometrics Special Interest Group, BIOSIG, pp. 1–5. IEEE, (2018)
74. Alonso-Fernandez, F., et al.: Cross-sensor periocular biometrics: a comparative benchmark including smartphone authentication. CoRR (2019). abs/1902.08123. http://arxiv.org/abs/1902.08123
75. Howard, A.G., et al.: Mobilenets: efficient convolutional neural networks for mobile vision applications. CoRR (2017). http://arxiv.org/abs/1704.04861

76. He, K., et al.: Identity mappings in deep residual networks. CoRR (2016). http://arxiv.org/abs/1603.05027
77. Zhang, K. et al.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process. Lett 23(10), 1499–1503 (2016)
78. Vapnik, V.N.: The nature of statistical learning theory. Springer-Verlag New York, Inc., New York (1995)
79. van der Maaten, L., Hinton, G.: Visualising data using t-sne. J. Mach. Learn. Res. 9(86), 2579–2605 (2008). http://jmlr.org/papers/v9/vandermaaten08a.html
80. Levi, G., Hassner, T.: Age and gender classification using convolutional neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 34–42. IEEE (2015)
81. Parkhi, O.M., et al.: Deep face recognition. In: Xie, M.W.J., Tam, G.K.L. (eds.) Proceedings of the British Machine Vision Conference (BMVC), pp. 411–4112. BMVA Press (2015)
82. Morales, A., et al.: SensitiveNets: learning agnostic representations with application to face images, IEEE Trans. Pattern. Anal. Mach. Intell. 43(6), 2158–2164 (2020)
83. Rothe, R., Timofte, R., Van Gool, L.: Deep expectation of real and apparent age from a single image without facial landmarks. Int. J. Comput. Vis. 126, 144–157 (2018)

How to cite this article: Alonso-Fernandez, F., et al.: Facial masks and soft-biometrics: Leveraging face recognition CNNs for age and gender prediction on mobile ocular images. IET Biom. 10(5), 562–580 (2021). https://doi.org/10.1049/bme2.12046