Facial Soft Biometrics Detection on Low Power Devices

Manolis Vasileiadis, Georgios Stavropoulos and Dimitrios Tzovaras
Information Technologies Institute, Centre for Research and Technology Hellas
6th km Charilaou-Thermi Rd, Thessaloniki, 57001 Greece
{mavasile, stavrop, dimitrios.tzovaras}@iti.gr

Abstract

Soft biometric traits have been proven to enhance person identification accuracy, when used complementary to primary biometric traits. They present a series of advantages such as compliance to the human language, robustness to low quality data, non-intrusive and consent free acquisition, and privacy preservation, increasing their applicability in realistic conditions. They can be extracted from a variety of individual modalities, with the human face being considered as the most informative source of attributes, as it provides rich geometrical and texture features. Recent advances in computer vision have allowed the accurate detection of such features under varying, non-ideal capturing conditions, with this increase in detection capacity, however, coming at the cost of high computational complexity. Meanwhile, the research and market interest has shifted towards the implementation of such methods on low power devices (i.e mobile phones), with data security concerns favoring on-device offline computation instead of cloud-based services. Towards this end, and taking into consideration recent advances in computationally efficient CNN design and multitask learning, we propose a novel CNN architecture, suitable for real time implementation on low power devices, which simultaneously performs gender, age, race, eyes state, eyewear, smile, beard and moustache estimation from unconstrained face images. The architecture employs the Mobilenet architecture and exploits the correlation between the individual biometric features, performing comparably to three state-of-the-art face analysis systems, while requiring significantly lower computational resources.

1. Introduction

Soft biometrics can be defined as non-unique personal attributes of physical, behavioral or material nature, which can be used for person description and identity verification. They are typically gleaned from primary biometric data, in automated fashion, and classify people in predefined human-language interpretable categories [64, 15, 46]. While they do not present the discriminative capacity to allow accurate person identification on their own [3], soft biometric features have been proven to increase person identification accuracy in unconstrained settings, when used complementary to primary biometric traits (also known as hard biometrics) [71, 79, 28], leading to several research approaches towards the fusion of soft and hard biometric features [93, 66, 2, 68, 26].

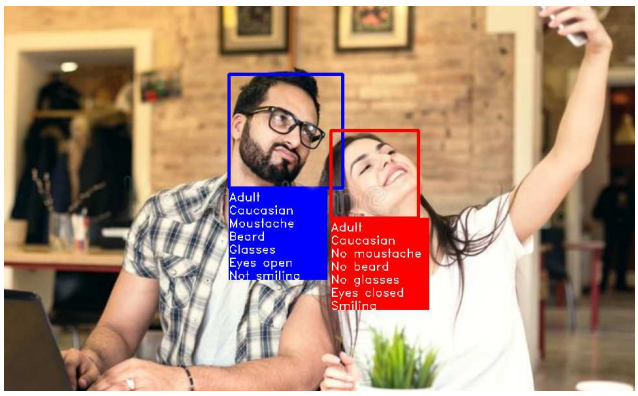

Figure 1. The proposed methodology utilizes a computationally efficient deep network architecture to simultaneously extract 8 facial soft biometric features from unconstrained faces images

In addition to their complementary utilization towards person identification, soft biometrics further present multiple benefits, including:

- *Compliance to human language:* Soft biometrics provide a description that can be interpreted and easily understood by humans (i.e. "tall, blonde, female") making them suitable for scenarios where only a verbal description is available (i.e. eyewitness statement), such as surveillance or police reports [73, 71].

- *Robustness to challenging conditions:* The lower discriminative capacity of soft biometrics allows their deduction from data of low quality, as they present robustness to viewpoint variance, illumination, occlusions and low resolution [45].

- *Non-intrusive, consent-free:* Their robustness to low quality data, allows the extraction of soft biometrics in unconstrained settings (i.e. from long distance), thus not requiring the consent or cooperation of the subject, rendering the data acquisition process almost imperceptible.

- *Privacy preservation:* Soft biometric traits are non-unique personal attributes, which provide only a partial description of a person's appearance and behavior, thus allowing the preservation of his identity. This feature can be of high significance when it comes to the capture and storage of such data, especially in the light of recent strict personal data protection regulations (i.e EU GDPR).

The human face is considered as the most informative source of attributes, as the facial features provide a rich and highly discriminative representation of the human appearance, allowing the extraction of a multitude of soft biometric traits (i.e. gender, age, race, facial hair, eyewear etc.), with the field of person recognition from facial soft biometrics having garnered a significant amount of attention from the research community [49, 80, 93, 5, 4, 28]. Even though facial soft biometrics are relatively robust to low quality data, their extraction in unconstrained settings ("in the wild") can still be a challenging process, as the large variability in capturing conditions (viewpoint, illumination etc.) as well as the diversity of similar facial features among people of different race [19], can significantly impact the prediction accuracy. While initial approaches would undertake the extraction of only a single soft biometric trait, focusing mainly on age [35, 18, 65], gender [59, 71] and race [22], not leveraging the potential correlation between those features, recent works have shown that jointly learning correlated tasks can improve the overall performance of each individual task [97, 9, 69, 7].

This increase in detection capacity, however, has come at the cost of computational complexity, as high complexity detection models are usually employed [69, 7, 31], requiring high performance hardware to perform in timely fashion. Meanwhile, the research and market interest has shifted towards the implementation of such methods on low power devices (i.e mobile phones). While cloud-based services offer an indirect solution to this need, executing the complex calculations remotely in high performance workstations, the increased data transmission latency, and more importantly the significant security concerns in regards to the transmission, Over-the-Air, of potentially sensitive personal data to third-party vendors, have reinforced the requirement for offline, on-device computation.

Towards this end, and taking into consideration recent advances in computationally efficient CNN design and Multitask learning, we propose a novel CNN architecture, suitable for real time implementation on low power devices, which simultaneously performs gender, age, race, eyes state, eyewear, smile, beard and moustache estimation from unconstrained face images. The proposed method employs the Mobilenet [40] architecture, along with the TensorflowLite quantization scheme [44], and exploits the correlation between the individual biometric features, performing comparably to three state-of-the-art face analysis systems, while requiring significantly lower computational resources. The main contributions of this paper are:

- A deep CNN architecture that simultaneously performs gender, age, ethnicity, eyes state, eyewear, smile, beard and moustache estimation from unconstrained face images, taking advantage of the correlation of each of these tasks

- The architecture employs a computationally efficient and parameterizable design, significantly reducing the computational complexity, compared to other SOA methods, thus offering real-time performance in low power devices

- Through experimental evaluation on a publicly available facial soft biometrics dataset, the proposed methodology achieves very high recognition rates, comparable to three state-of-the-art face analysis systems

The rest of the paper is organized as follows. Section 2 provides a summary of the state-of-the-art in the field of facial soft biometrics detection and efficient network design. Section 3 describes the proposed network architecture. Section 4 presents the results from the experimental evaluation and the computational complexity estimation and Section 5 concludes the paper.

2. Related Work

2.1. Individual Facial Soft Biometrics

Early facial soft biometric-based approaches undertook the extraction of individual soft biometric traits, with the majority of the research work focusing on age [35, 18, 65], gender [59, 71] and race [22] estimation from face images.

Towards age estimation, a variety of face representation models have been utilized including: *Wrinkle Models* [37], *Active Appearence Models (AAM)* [51], *Aging Pattern Subspace* [25], *Age Manifolds* [32] and *Biologically Inspired Models* [34]. Moreover, general-purpose image features have also been employed, such as *Local Binary Patterns (LBPs)* [30] and *Gabor Features* [23]. Meanwhile, in recent years facial age estimation approaches have moved towards the utilization of deep network architectures, with a large number of CNN architectures proposed for age estimation and grouping from a single face image [57, 63, 16, 11, 55]

Towards gender classification, early attempts employed networks such as *ANNs* [27] and *HyberBF* [67], while subsequent approaches used global image features utilizing *SVM* [14] and *AdaBoost* [6] classifiers. The same classifiers were also later used in conjunction with local image features including *LBPs* [75], *SIFT features* [81] and *Haar wavelets* [90]. Meanwhile the use of *Lookup Tables* was proposed in [84].

Towards race estimation (also referred in literature as ethnicity), similar to the above mentioned traits, a multitude of methods have been proposed, utilizing global color features [88] and local image features, such as *LBPs* [72], *Gabor features* [39], *Haar wavelets* [85]. Moreover, as in the case of age estimation, more recent research efforts have been mainly focused towards the use of deep CNNs, combining convolutional structures for image feature extraction and fully-connected layers for the final race classification [83, 96, 38].

Similar efforts, albeit more limited, have been carried out towards the detection of other facial traits, utilizing, in a similar manner to the methods mentioned above, feature / classifier combinations, as well deep CNNs, for eyewear [86, 20, 17], smile [76, 24, 10], facial hair [62, 53, 52] and eyes state [92, 94, 29] detection.

2.2. Joint Facial Feature Learning

While the individual facial feature extraction methods, described above, have achieved high accuracy results, recent research efforts have shown that jointly learning correlated tasks, known as Multitask learning [8], can improve the overall performance of each individual task. Zhu and Ramanan [97] presented a unified model for face detection, pose estimation, and landmark estimation in real-world, cluttered images, using a model based on a mixture of trees with a shared pool of parts. Chen et al. [9] combined face alignment with detection, learning the two tasks jointly in the same cascade framework, while in [33] a framework for joint estimation of age, gender and ethnicity was proposed, exploring *Canonical Correlation Analysis* and *Partial Least Square* based models. Meanwhile, Eidinger et al. [18] extracted *LBP* features to train *linear-SVM* models for age and gender estimation in unconstrained face images.

Moving towards *CNN* architectures, Levi and Hassner [54] were the first to utilize them for simultaneous age and gender classification. In [69] the *Hyperface* model was introduced, a deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition, with [70] extending it towards more extensive face analysis. Gunther at al. [31] present an alignment-free facial attribute classification technique, while Cao et al. [7] integrate *Partially Shared* (PS) structures and local constraints together to help the framework learn better attribute representations.

2.3. Computationally Efficient CNNs

Multiple recent research efforts have focused in building small and efficient neural networks suitable for systems with limited resources, such as mobile devices. A common approach is to reduce the number of parameters in the convolutions, with the *MobileNet* [40, 74], *Shufflenet* [95, 58] and *Xception*[13] models utilizing depth-wise separable convolutions [77]. An alternative approach was proposed in [82], introducing factorized convolutions, while Jin et al. [47] proposed the use of topological connections for further reducing computational requirements (*Flattened networks*). Other small networks include the *Squeezenet* [42] which used a bottleneck approach to design a very small network, *structured transform networks* [78] and *deep fried convnets* [91].

Instead of reducing the convolution parameters before training, a different approach is to obtain a small network by shrinking, factorizing or compressing pre-trained networks. The most popular techniques for compressing pre-trained networks are 1) quantization [87, 44], in which filter kernels and weighting matrices are quantized, 2) hashing [12], which uses a low-cost hash function to randomly group connection weights into hash buckets, and all connections within the same hash bucket share a single parameter value and pruning, and 3) Huffman coding [36] that is able to further reduce the size of the networks using Huffman coding on the weights of the network.

3. Proposed Methodology

The proposed methodology follows the *MobileNet* [40] deep network architecture, combined with a fully connected multi-part classification layer, in order to jointly estimate 8 facial soft biometric traits, described in Table 1 below, from unconstrained RGB face images.

Table 1. The facial soft biometric traits extracted by the proposed method

Gender	male / female
Age Group	infant / child / young adult / adult / senior
Race	caucasian / african / asian / hindi / mixed
Eyes State	open / partially open / closed
Eyewear	glasses / no glasses
Smile	smiling / not smiling
Beard	beard / no beard
Moustache	moustache / no moustache

3.1. Depthwise Separable Convolutions

The Mobilenet network architecture utilizes a highly efficient convolutional structure, called *Depthwise Separable*

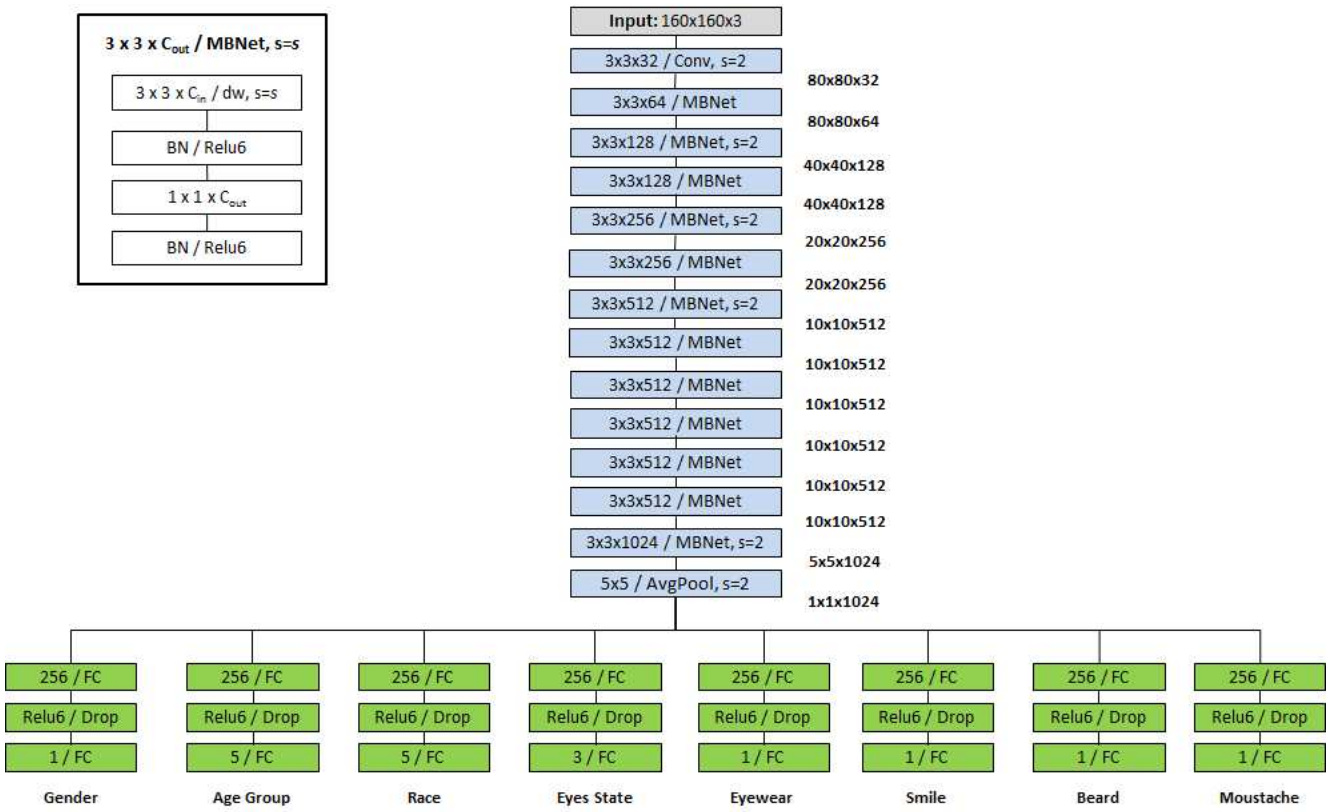

Figure 2. The proposed network architecture: The *feature extraction* stage (light blue) generates the low level image features from a $160 \times 160px$ face image (light gray). It employs a truncated version of the *MobileNet* [40] network architecture, composed of multiple stacked depthwise weparable convolution layers. The output of the final convolutional layer is passed through an Average Pooling layer and is fed to the *classification* stage (green), which encodes the classification-specific information for the detection of facial soft biometric traits, using an individual two-fully-connected-layer structure for each of the target facial soft biometrics

Convolution [77] (Figure 2) in order to reduce the computational complexity of standard convolutional layers, without affecting their representational capacity. Depthwise Separable Convolutions are composed by an computationaly efficient 3×3 depthwise convolution, which applies a single filter to each input channel, followed by a standard 1×1 pointwise convolution which combines the outputs of the former, with Batch Normalization[43] and Relu6 [50] activation applied after each layer.

While a standard convolution both filters and combines inputs into a new set of outputs in a single step, the depthwise separable convolution splits this process into two steps, with this factorization significantly reducing computation cost and model size.

3.2. Network Architecture

The network architecture, presented in detail in Figure 2, is split in two stages: *feature extraction* and *classification*.

The *feature extraction* stage generates the low level image features from the input. It employs a truncated version of the Mobilenet [40] network architecture, using only the fully-convolutional layers, since the final fully-connected layers are used for specific-object classification and thus are discarded. More specifically, the feature extraction stage is comprised of an initial dense $3 \times 3 \times 32$ layer, followed by 12 depthwise separable convolutional layers which gradually reduce the spatial dimensions of the feature maps while increasing their depth. The output of the final convolutional layer passes through an average pooling layer in order to reduce the spatial dimensions of the feature maps to 1×1, transforming them into a 1024 dimensional feature vector.

The *classification* stage encodes the classification-specific information for the detection of facial soft biometric traits. It receives as input the feature vector generated by the feature extraction stage and passes it to 8 individual two-layer classification structures, one for each of the target facial soft biometrics. Each structure is comprised of two fully connected layers: a 256 neuron layer which reduces the dimensionality of the input feature vector, followed by a second layer which produces the final probability.

3.3. Computational Complexity Configuration

In order to easily manipulate the trade-off between computational complexity and accuracy, a width multiplier hyperparameter $a \in (0, 1]$ is introduced. The width multiplier is applied to all the convolutional and fully connected layers, uniformly reducing the width of the network at each layer. The baseline configuration described above corresponds to $a = 1$, while typical values of 0.75, 0.50 and 0.25 are commonly used to define smaller models.

Moreover, post-training quantization [44] is employed, significantly reducing the computational cost by substituting floating point operations with inexpensive fixed point arithmetic. While training is performed in floating point accuracy, during integer-arithmetic-only inference, the input and output are represented as 8-bit integers according to an affine mapping of integers q to real numbers r:

$$r = S\,(q - Z) \tag{1}$$

where S and Z the quantization parameters, with $S > 0$ an arbitrary real scaling factor, and Z an integer corresponding to the real value 0. The quantization scheme uses a single set of quantization parameters for all values within each activations array and within each weights array; separate arrays use separate quantization parameters.

3.4. Training

Loss function During the training phase, the individual loss L_f is estimated for each of the eight target facial biometric traits f:

- for the binary biometric features $b \in \{gender, eyewear, smile, beard, moustache\}$ the binary cross entropy loss is computed:

$$L_b = -\,(1 - b) \cdot \log\,(1 - p_b) - b \cdot \log\,(p_b) \tag{2}$$

  where $b = 0, 1$ the two possible labels and p_b the probability that the input is assigned to label 1.

- for the multidimensional features $m \in \{age\,group, race, eyes\,state\}$ the softmax multinomial cross entropy loss is employed:

$$L_m = -\sum_{k=1}^{K} m_k \log\,(p_{mk}) \tag{3}$$

  where $m_k = 1$ if the input belongs to class k and p_{mk} is the probability that the input is assigned to class k.

The overall loss L is defined as the sum of the individual losses of each facial biometric feature:

$$L = \sum_{b} L_b + \sum_{m} L_m \tag{4}$$

Optimization For the optimization of the objective function, the Adam [48] gradient-based optimization algorithm is employed, with learning rate $a = 0.0004$ and momentum $b = 0.9$

4. Experimental Evaluation

The proposed methodology is trained and experimentally evaluated on the *LFW Soft Biometrics* dataset [28]. The dataset is an extension of the *Labelled Faces in the Wild (LFW)* [41] dataset, and includes manually annotated groundtruth values for 11 facial attributes (*gender, age, race, glasses, beard moustache, forehead, mouth, eyes, smiling and pose*) for 13233 unconstrained face images of 5749 people, sized $250 \times 250px$.

Since the authors do not provide a train/test split, the dataset is randomly divided into four non-overlapping subsets, and is trained and evaluated using the leave-one-out strategy: one subset is used as the test set, while the other three are used as the training set. This process is performed once for each subset with the average prediction accuracy calculated for every one of the eight target facial biometric traits. Each training session is performed for 50 epochs, using a batch size of 32 samples, with the feature extraction stage initialized from the pre-trained Mobilenet model.

4.1. Biometrics Detection Accuracy

Table 2 presents the accuracy results from the experimental evaluation of multiple configurations of the proposed methodology on the 8 target facial soft biometric traits of the *LFW Soft Biometrics* dataset. The overall accuracy score of all the features for the baseline model ($a = 1$) is 92.1%, showcasing the architecture's potential towards accurate facial biometric detection.The proposed method achieves very high accuracy scores for the *gender, eyes state, eyewear, beard* and *moustache* features, exceeding 94% accuracy, while the *race* and *smile* features surpass the 89% accuracy threshold. *Age group* estimation, on the other hand, presents a lower accuracy score of 75.6%, which is mainly attributed to large overlap between the *young adult / adult* and *adult / senior* age groups.

For smaller width multiplier values ($a < 1$), the results follow a similar pattern, with the $a = 0.75$ and $a = 0.50$ models presenting an average reduction in accuracy of less than 1% compared to the baseline model, while even the extremely narrow $a = 0.25$ model achieves a mean accuracy score of 90%. Furthermore, the quantized models also report high accuracy results, presenting, on average, just a $\sim 1\%$ drop in overall accuracy, compared to the corresponding floating point models.

Moreover, the proposed method is comparatively evaluated to three state-of-the-art face analysis systems, *Hyperface* [69], *Face++* [60] and *Microsoft Cognitive Services*

Table 2. Comparative experimental evaluation of multiple configurations of the proposed methodology and three state-of-the-art face analysis methods, on the 8 target facial soft biometric traits of the LFW Soft Biometrics dataset. Accuracy scores for *Hyperface* were extracted by retraining the Hyperface-Alexnet architecture on the dataset, while the results for *Face++* and *MS Cognitive services* were taken from [28]

Model	**Gender**	**Age**	**Race**	**Eyes state**	**Eyewear**	**Smile**	**Beard**	**Moustache**	**MEAN**
a=1.00	0.967	0.756	0.897	0.941	0.980	0.891	0.979	0.961	0.921
a=0.75	0.957	0.752	0.893	0.915	0.979	0.871	0.975	0.962	0.913
a=0.50	0.961	0.740	0.886	0.936	0.974	0.884	0.976	0.967	0.915
a=0.25	0.958	0.728	0.858	0.898	0.963	0.876	0.963	0.956	0.900
a=1.00 quan	0.968	0.745	0.892	0.942	0.979	0.891	0.936	0.928	0.910
a=0.75 quan	0.958	0.751	0.890	0.918	0.978	0.874	0.935	0.935	0.905
a=0.50 quan	0.962	0.736	0.886	0.937	0.973	0.884	0.935	0.931	0.906
a=0.25 quan	0.959	0.726	0.855	0.904	0.958	0.878	0.936	0.919	0.892
Hyperface [69]	0.979	0.772	0.901	0.936	0.980	0.902	0.981	0.970	0.928
Face++ [60]	0.911	0.388	0.874	-	0.922	-	-	-	-
MS Cognitive Services [61]	0.929	0.593	-	-	0.917	-	0.938	0.941	-

[61]. The baseline model almost matches the overall accuracy of Hyperface, while even the smallest $a = 0.25$ quantized model performs comparably, with an average drop in accuracy of 3.5%. Meanwhile, all model configurations outperform Face++ and MS Cognitive Services on almost all target facial soft biometrics [1].

Finally, some qualitative results from the experimental evaluation on the *LFW Soft Biometrics* dataset, including both successful and failed examples, are presented in Figure 4. Additionally, some examples from the *NIST SD32 - MEDS* [21] and *Adience* [18] datasets are also shown.

4.2. Computational Complexity Evaluation

While the proposed methodology performs comparably to SOA face analysis methods, one of the main goals of this work is the generation of a model of low computational complexity, suitable for implementation in low power devices, thus making it necessary to establish an absolute metric that will describe the computational cost of an architecture and provide an estimation about its runtime. While actual runtime measurements could be used as such a metric, they cannot be considered an absolute metric as their values depend on the platform upon which the algorithm is benchmarked, presenting significant variations between different hardware and software configurations.

Meanwhile, the task in hand requires framework agnostic metrics. Towards this end, the following metrics are introduced (Table 3):

- **Multiply-Accumulate Operations (MACs)**: MACs describe the number of arithmetic computations required to perform a function. In the case of neural networks most of the computations are dot products (multiplication followed by addition), with a single multiplication-addition corresponding to 1 MAC
- **Memory Access Operations (MEMs)**: MEMs describe the amount of memory access operations required in order to read all the data necessary for the computation and save its result. Following, for simplicity, a naive approach, where each value has to be read and written every time (discarding advanced techniques such as caching etc.), for each layer there are three memory access operation groups: a) read the input data, b) read the filter weights, c) write the output.
- **Network Parameters**: The size of a network is calculated by counting the number of trainable variables at each layer. It not only affects the computational cost (more parameters equal more MACs and MEMs) of a network, but also dictates the minimum amount of storage required to save it

Table 3. Estimation of the framework agnostic computational complexity metrics for fully connected and convolutional layers (biases skipped for simplicity)

Layer Type	**Fully Connected**	**Convolutional**
Input Dims	K	$H \times W \times C_{in}$
Layer	n neurons	$k \times k \times C_{out}$ stride= s groups= g
MACs	$K \cdot n$	$\frac{H \cdot W \cdot C_{in} \cdot k^2 \cdot C_{out}}{g \cdot s^2}$
MEMs	$K + K \cdot n + n$	$H \cdot W \cdot C_{in} +$ $C_{in} \cdot k^2 \cdot C_{out}/g +$ $H \cdot W \cdot C_{out}/s^2$
Params	$K \cdot n$	$C_{in} \cdot k^2 \cdot C_{out}/g$

[1] accuracy scores for *Hyperface* were extracted by retraining the Hyperface-Alexnet architecture on the LFW Soft Biometrics dataset, while the results for *Face++* and *MS Cognitive services* were taken from [28]

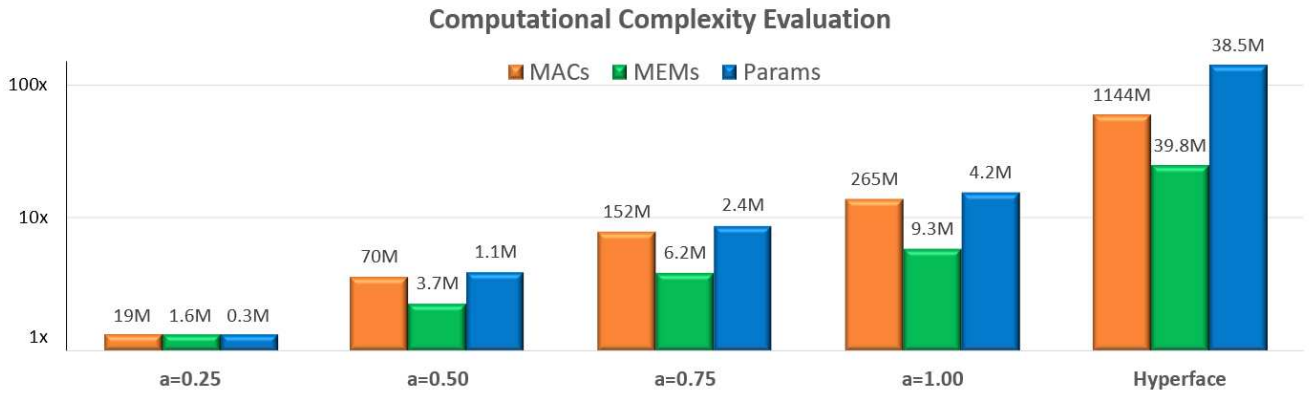

Figure 3. Estimation of the MACs, MEMs and Network Parameters for multiple configurations of the proposed method and the Hyperface-Alexnet model [69]. The bars correspond to the relative increase of the metrics for each model relevant to the $a = 0.25$ model (logarithmic scale), while the data labels present the absolute values for each metric. Only the cost of the convolutional and fully connected layers was estimated, as the rest of the layers (pooling, activation, normalization etc.) do not significantly impact the overall values.

Based on the above metrics, the computational complexity of the proposed method, for multiple width multiplier values, is estimated and compared with the state-of-the-art Hyperface-Alexnet [69] model (Figure 3). From the results, it becomes evident that the proposed methodology presents a significant reduction in computational complexity, with the baseline model $a = 1.00$ requiring approximately $4.3\times$ less MACs and MEMs, and $9\times$ less network parameters than Hyperface. Moreover, when compared to the narrow $a = 0.25$ model the gains in computational complexity are even more extreme with a massive $60\times$ reduction in MACs, $25\times$ in MEMs and $128\times$ in Network parameters. These results, combined with high accuracy scores presented in Section 4.1, clearly showcase the capacity of the proposed method for accurate and computationally efficient facial soft biometrics detection, suitable for implementation on low power devices.

4.3. Inference

The proposed algorithm is implemented in Python, utilizing the Tensorflow/Lite deep learning framework [1] for training the network and deploying the final inference models. Since the proposed method requires as input only facial images, a face detector needs to be employed to provide the face regions. Staying in line with the computationally efficient design, the quantized Mobilenet-SSD [56] detector is used, trained on the WIDERFACE [89] face detection dataset.

In order to better showcase the real-life performance of the proposed method, the final inference models are benchmarked on two mobile devices based on ARM platforms. The baseline model takes $90ms$ to process an image on a single mid-range Qualcomm Snapdragon 636 processor, and $178ms$ on a older HiSilicon Kirin 650 processor. Reducing the width multiplier by 0.25 approximately doubles the inference speed on both devices, while moving to the quantized versions further reduces the latency by a factor of $2\times$, with the smallest quantized $a = 0.25$ model requiring just $7ms$ to process an image, and only $300KB$ in storage space. The detailed benchmark results are presented in Table 4 below.

Table 4. Actual inference time and model size on two mobile devices (a smaller number represents better performance). The platforms are based on single Qualcomm Snapdragon 636 and a HiSilicon Kirin 650 processors

Model	**Inference Time (ms)**		**Model**
	SDM 636	**Kirin 650**	**Size (MB)**
Mbnet SSD quan	75	60	3.1
a=1.00	90	178	17
a=0.75	50	110	9.6
a=0.50	25	65	4.3
a=0.25	13	28	1.1
a=1.00 quan	45	70	4.3
a=0.75 quan	30	48	2.4
a=0.50 quan	17	25	1.1
a=0.25 quan	7	11	0.3

5. Conclusion

This paper introduced a computationally efficient CNN architecture for the simultaneous estimation of 8 facial soft biometric features from unconstrained face images: *gender, age, race, eyes state, eyewear, smile, beard* and *moustache*, suitable for deployment on low power devices. The proposed method employs the Mobilenet architecture along with a post-training quantization scheme, exploiting the correlation between the individual biometric features, and achieves very high accuracy results on a publicly available

Figure 4. Qualitative results from the experimental evaluation of the baseline model on the *LFW Soft Biometrics* (first and second rows), *Adience* (third row) and *NIST SD32 - MEDS* (fourth row) datasets . Facial features depicted in ***bold italics*** indicate wrong detections. The facial regions were cropped using the quantized Mobilenet-SSD face detector.

facial soft biometrics dataset. The architecture matches the performance of state-of-the-art face analysis models, while requiring significantly lower computational resources, with benchmarking on two low power mobile platforms showcasing its applicability in real life conditions.

Acknowledgments

This work is co-funded by the European Union (EU) within the SMILE project under grant agreement number 740931. The SMILE project is part of the EU Framework Program for Research and Innovation Horizon 2020

References

[1] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, pages 265–283, 2016. 7

[2] M. C. D. C. Abreu and M. Fairhurst. Enhancing identity prediction using a novel approach to combining hard-and soft-biometric information. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 41(5):599–607, 2011. 1

[3] Z. Akhtar, A. Hadid, M. Nixon, M. Tistarelli, J.-L. Dugelay, and S. Marcel. Biometrics: In search of identity and security (q & a). *IEEE MultiMedia*, 2017. 1

[4] N. Y. Almudhahka, M. Nixon, and J. Hare. Comparative face soft biometrics for human identification. In *Surveillance in Action*, pages 25–50. Springer, 2018. 2

[5] N. Y. Almudhahka, M. Nixon, and J. S. Hare. Unconstrained human identification using comparative facial soft biometrics. In *IEEE International Conference on Biometrics Theory, Applications and Systems*, pages 1–6. IEEE, 2016. 2

[6] S. Baluja and H. A. Rowley. Boosting sex identification performance. *International Journal of computer vision*, 71(1):111–119, 2007. 3

[7] J. Cao, Y. Li, and Z. Zhang. Partially shared multi-task convolutional neural network with local constraint for face attribute learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4290–4299, 2018. 2, 3

[8] R. Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997. 3

[9] D. Chen, S. Ren, Y. Wei, X. Cao, and J. Sun. Joint cascade face detection and alignment. In *European Conference on Computer Vision*, pages 109–122. Springer, 2014. 2, 3

[10] J. Chen, Q. Ou, Z. Chi, and H. Fu. Smile detection in the wild with deep convolutional neural networks. *Machine vision and applications*, 28(1-2):173–183, 2017. 3

[11] S. Chen, C. Zhang, M. Dong, J. Le, and M. Rao. Using ranking-cnn for age estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[12] W. Chen, J. Wilson, S. Tyree, K. Weinberger, and Y. Chen. Compressing neural networks with the hashing trick. In *International Conference on Machine Learning*, pages 2285–2294, 2015. 3

[13] F. Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017. 3

[14] N. Costen, M. Brown, and S. Akamatsu. Sparse models for gender classification. In *Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on*, pages 201–206. IEEE, 2004. 3

[15] A. Dantcheva, P. Elia, and A. Ross. What else does your biometric data reveal? a survey on soft biometrics. *IEEE Transactions on Information Forensics and Security*, 11(3):441–467, 2016. 1

[16] Y. Dong, Y. Liu, and S. Lian. Automatic age estimation based on deep learning algorithm. *Neurocomputing*, 187:4–10, 2016. 2

[17] S. Du, J. Liu, Y. Liu, X. Zhang, and J. Xue. Precise glasses detection algorithm for face with in-plane rotation. *Multimedia Systems*, 23(3):293–302, 2017. 3

[18] E. Eidinger, R. Enbar, and T. Hassner. Age and gender estimation of unfiltered faces. *IEEE Transactions on Information Forensics and Security*, 9(12):2170–2179, 2014. 2, 3, 6

[19] G. Farinella and J.-L. Dugelay. Demographic classification: Do gender and ethnicity affect each other? In *Informatics, Electronics & Vision (ICIEV), 2012 International Conference on*, pages 383–390. IEEE, 2012. 2

[20] A. Fernández, R. Casado, and R. Usamentiaga. A real-time big data architecture for glasses detection using computer vision techniques. In *Future Internet of Things and Cloud (FiCloud), 2015 3rd International Conference on*, pages 591–596. IEEE, 2015. 3

[21] A. P. Founds, N. Orlans, W. Genevieve, and C. I. Watson. Nist special database 32-multiple encounter dataset ii (meds-ii). Technical report, 2011. 6

[22] S. Fu, H. He, and Z.-G. Hou. Learning race from face: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 36(12):2483–2509, 2014. 2

[23] F. Gao and H. Ai. Face age classification on consumer images with gabor feature and fuzzy lda method. In *International Conference on Biometrics*, pages 132–141. Springer, 2009. 2

[24] Y. Gao, H. Liu, P. Wu, and C. Wang. A new descriptor of gradients self-similarity for smile detection in unconstrained scenarios. *Neurocomputing*, 174:1077–1086, 2016. 3

[25] X. Geng, Z.-H. Zhou, and K. Smith-Miles. Automatic age estimation based on facial aging patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 29(12):2234–2240, 2007. 2

[26] A. E. K. Ghalleb and N. E. B. Amara. Soft and hard biometrics for the authentication of remote people in front and side views. *International Journal of Applied Engineering Research*, 11(14):8120–8127, 2016. 1

[27] B. A. Golomb, D. T. Lawrence, and T. J. Sejnowski. Sexnet: A neural network identifies sex from human faces. In *NIPS*, volume 1, page 2, 1990. 3

[28] E. Gonzalez-Sosa, J. Fierrez, R. Vera-Rodriguez, and F. Alonso-Fernandez. Facial soft biometrics for recognition in the wild: Recent works, annotation, and cots evaluation. *IEEE Transactions on Information Forensics and Security*, 13(8):2001–2014, 2018. 1, 2, 5, 6

[29] C. Gou, Y. Wu, K. Wang, F.-Y. Wang, and Q. Ji. Learning-by-synthesis for accurate eye detection. In *Pattern Recognition (ICPR), 2016 23rd International Conference on*, pages 3362–3367. IEEE, 2016. 3

[30] A. Gunay and V. V. Nabiyev. Automatic age classification with lbp. In *Computer and Information Sciences, 2008. ISCIS'08. 23rd International Symposium on*, pages 1–4. IEEE, 2008. 2

[31] M. Günther, A. Rozsa, and T. E. Boult. Affact: Alignment-free facial attribute classification technique. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pages 90–99. IEEE, 2017. 2, 3

[32] G. Guo, Y. Fu, T. S. Huang, and C. R. Dyer. Locally adjusted robust regression for human age estimation. In *Applications of Computer Vision, 2008. WACV 2008. IEEE Workshop on*, pages 1–6. IEEE, 2008. 2

[33] G. Guo and G. Mu. A framework for joint estimation of age, gender and ethnicity on a large database. *Image and Vision Computing*, 32(10):761–770, 2014. 3

[34] G. Guo, G. Mu, Y. Fu, and T. S. Huang. Human age estimation using bio-inspired features. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 112–119. IEEE, 2009. 2

[35] H. Han, C. Otto, and A. K. Jain. Age estimation from face images: Human vs. machine performance. In *IEEE International Conference on Biometrics*, pages 1–8. IEEE, 2013. 2

[36] S. Han, H. Mao, and W. J. Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*, 2015. 3

[37] J. Hayashi, M. Yasumoto, H. Ito, and H. Koshimizu. Method for estimating and modeling age and gender using facial image processing. In *Virtual Systems and Multimedia, 2001. Proceedings. Seventh International Conference on*, pages 439–448. IEEE, 2001. 2

[38] Z. Heng, M. Dipu, and K.-H. Yap. Hybrid supervised deep learning for ethnicity classification using face images. In *Circuits and Systems (ISCAS), 2018 IEEE International Symposium on*, pages 1–5. IEEE, 2018. 3

[39] S. Hosoi, E. Takikawa, and M. Kawade. Ethnicity estimation with facial images. In *Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on*, pages 195–200. IEEE, 2004. 3

[40] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 2, 3, 4

[41] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007. 5

[42] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016. 3

[43] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015. 4

[44] B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, and D. Kalenichenko. Quantization and training

of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2704–2713, 2018. 2, 3, 5

[45] A. K. Jain, S. C. Dass, and K. Nandakumar. Soft biometric traits for personal recognition systems. In *Biometric Authentication*, pages 731–738. Springer, 2004. 1

[46] A. K. Jain, K. Nandakumar, and A. Ross. 50 years of biometric research: Accomplishments, challenges, and opportunities. *Pattern Recognition Letters*, 79:80–105, 2016. 1

[47] J. Jin, A. Dundar, and E. Culurciello. Flattened convolutional neural networks for feedforward acceleration. *arXiv preprint arXiv:1412.5474*, 2014. 3

[48] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. 5

[49] B. F. Klare, S. Klum, J. C. Klontz, E. Taborsky, T. Akgul, and A. K. Jain. Suspect identification based on descriptive facial attributes. In *IEEE International Joint Conference on Biometrics*, pages 1–8. IEEE, 2014. 2

[50] A. Krizhevsky and G. Hinton. Convolutional deep belief networks on cifar-10. *Unpublished manuscript*, 40(7), 2010. 4

[51] A. Lanitis, C. Draganova, and C. Christodoulou. Comparing different classifiers for automatic age estimation. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 34(1):621–628, 2004. 2

[52] T. H. N. Le, K. Luu, and M. Savvides. Fast and robust self-training beard/moustache detection and segmentation. In *Biometrics (ICB), 2015 International Conference on*, pages 507–512. IEEE, 2015. 3

[53] T. H. N. Le, K. Luu, K. Seshadri, and M. Savvides. Beard and mustache segmentation using sparse classifiers on self-quotient images. In *Image Processing (ICIP), 2012 19th IEEE International Conference on*, pages 165–168. IEEE, 2012. 3

[54] G. Levi and T. Hassner. Age and gender classification using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 34–42, 2015. 3

[55] H. Liao, Y. Yan, W. Dai, and P. Fan. Age estimation of face images based on cnn and divide-and-rule strategy. *Mathematical Problems in Engineering*, 2018, 2018. 2

[56] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016. 7

[57] X. Liu, S. Li, M. Kan, J. Zhang, S. Wu, W. Liu, H. Han, S. Shan, and X. Chen. Agenet: Deeply learned regressor and classifier for robust apparent age estimation. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 16–24, 2015. 2

[58] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 116–131, 2018. 3

[59] E. Makinen and R. Raisamo. Evaluation of gender classification methods with automatically detected and aligned faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(3):541–547, 2008. 2

[60] Megvii Technology. Face++ cognitive services. https://www.faceplusplus.com. 5, 6

[61] Microsoft. Microsoft azure cognitive services. https://azure.microsoft.com/en-us/services/cognitive-services/. 6

[62] M. H. Nguyen, J.-F. Lalonde, A. A. Efros, and F. De la Torre. Image-based shaving. In *Computer Graphics Forum*, volume 27, pages 627–635. Wiley Online Library, 2008. 3

[63] Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua. Ordinal regression with multiple output cnn for age estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4920–4928, 2016. 2

[64] M. S. Nixon, P. L. Correia, K. Nasrollahi, T. B. Moeslund, A. Hadid, and M. Tistarelli. On soft biometrics. *Pattern Recognition Letters*, 68:218–230, 2015. 1

[65] G. Panis, A. Lanitis, N. Tsapatsoulis, and T. F. Cootes. Overview of research on facial ageing using the fg-net ageing database. *IET Biometrics*, 5(2):37–46, 2016. 2

[66] U. Park and A. K. Jain. Face matching and retrieval using soft biometrics. *IEEE Transactions on Information Forensics and Security*, 5(3):406–415, 2010. 1

[67] B. Poggio, R. Brunelli, and T. Poggio. Hyberbf networks for gender classification. 1992. 3

[68] A. Prakash and R. Mukesh. A biometric approach for continuous user authentication by fusing hard and soft traits. *IJ Network Security*, 16(1):65–70, 2014. 1

[69] R. Ranjan, V. M. Patel, and R. Chellappa. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 2, 3, 5, 6, 7

[70] R. Ranjan, S. Sankaranarayanan, C. D. Castillo, and R. Chellappa. An all-in-one convolutional neural network for face analysis. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 17–24. IEEE, 2017. 3

[71] D. A. Reid, S. Samangooei, C. Chen, M. S. Nixon, and A. Ross. Soft biometrics for surveillance: an overview. In *Handbook of statistics*, volume 31, pages 327–352. Elsevier, 2013. 1, 2

[72] S. H. Salah, H. Du, and N. Al-Jawad. Fusing local binary patterns with wavelet features for ethnicity identification. In *Proceedings of World Academy of Science, Engineering and Technology*, number 79, page 471. World Academy of Science, Engineering and Technology (WASET), 2013. 3

[73] S. Samangooei, B. Guo, and M. S. Nixon. The use of semantic human description as a soft biometric. In *IEEE International Conference on Biometrics: Theory, Applications and Systems*, pages 1–7. IEEE, 2008. 1

[74] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018. 3

[75] C. Shan. Gender classification on real-life faces. In *International Conference on Advanced Concepts for Intelligent Vision Systems*, pages 323–331. Springer, 2010. 3

[76] C. Shan. Smile detection by boosting pixel differences. *IEEE transactions on image processing*, 21(1):431–436, 2012. 3

[77] L. Sifre and S. Mallat. Rigid-motion scattering for image classification. *PhD thesis, Ph. D. thesis*, 1:3, 2014. 3, 4

[78] V. Sindhwani, T. Sainath, and S. Kumar. Structured transforms for small-footprint deep learning. In *Advances in Neural Information Processing Systems*, pages 3088–3096, 2015. 3

[79] P. Tome, J. Fierrez, R. Vera-Rodriguez, and M. S. Nixon. Soft biometrics and their application in person recognition at a distance. *IEEE Transactions on information forensics and security*, 9(3):464–475, 2014. 1

[80] P. Tome, R. Vera-Rodriguez, J. Fierrez, and J. Ortega-Garcia. Facial soft biometric features for forensic face recognition. *Forensic science international*, 257:271–284, 2015. 2

[81] J.-G. Wang, J. Li, W.-Y. Yau, and E. Sung. Boosting dense sift descriptors and shape contexts of face images for gender recognition. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 96–102. IEEE, 2010. 3

[82] M. Wang, B. Liu, and H. Foroosh. Factorized convolutional neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 545–553, 2017. 3

[83] W. Wang, F. He, and Q. Zhao. Facial ethnicity classification with deep convolutional neural networks. In *Chinese Conference on Biometric Recognition*, pages 176–185. Springer, 2016. 3

[84] B. Wu, H. Ai, and C. Huang. Lut-based adaboost for gender classification. In *International Conference on Audio-and Video-Based Biometric Person Authentication*, pages 104–110. Springer, 2003. 3

[85] B. Wu, H. Ai, and C. Huang. Facial image retrieval based on demographic classification. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 3, pages 914–917. IEEE, 2004. 3

[86] B. Wu, H. Ai, and R. Liu. Glasses detection by boosting simple wavelet features. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 1, pages 292–295. IEEE, 2004. 3

[87] J. Wu, C. Leng, Y. Wang, Q. Hu, and J. Cheng. Quantized convolutional neural networks for mobile devices. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4820–4828, 2016. 3

[88] Y. Xie, K. Luu, and M. Savvides. A robust approach to facial ethnicity classification on large scale face databases. In *Biometrics: Theory, Applications and Systems (BTAS), 2012 IEEE Fifth International Conference on*, pages 143–149. IEEE, 2012. 3

[89] S. Yang, P. Luo, C. C. Loy, and X. Tang. Wider face: A face detection benchmark. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 7

[90] Z. Yang, M. Li, and H. Ai. An experimental study on automatic face gender classification. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, volume 3, pages 1099–1102. IEEE, 2006. 3

[91] Z. Yang, M. Moczulski, M. Denil, N. de Freitas, A. Smola, L. Song, and Z. Wang. Deep fried convnets. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1476–1483, 2015. 3

[92] F. Yutian, H. Dexuan, and N. Pingqiang. A combined eye states identification method for detection of driver fatigue. 2009. 3

[93] H. Zhang, J. R. Beveridge, B. A. Draper, and P. J. Phillips. On the effectiveness of soft biometrics for increasing face verification rates. *Computer Vision and Image Understanding*, 137:50–62, 2015. 1, 2

[94] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling. Appearance-based gaze estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4511–4520, 2015. 3

[95] X. Zhang, X. Zhou, M. Lin, and J. Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6848–6856, 2018. 3

[96] Z. Zhao and A. Kumar. Accurate periocular recognition under less constrained environment using semantics-assisted convolutional neural network. *IEEE Transactions on Information Forensics and Security*, 12(5):1017–1030, 2017. 3

[97] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2879–2886. IEEE, 2012. 2, 3