DESCRIPTORS AND REGIONS OF INTEREST FUSION FOR GENDER CLASSIFICATION IN THE WILD. COMPARISON AND COMBINATION WITH CONVOLUTIONAL NEURAL NETWORKS

M. Castrillón-Santana, J. Lorenzo-Navarro and E. Ramón-Balmaseda

SIANI - Universidad de Las Palmas de Gran Canaria
Spain

ABSTRACT

Gender classification (GC) has achieved high accuracy in different experimental evaluations based mostly on inner facial details. However, these results do not generalize well in unrestricted datasets and particularly in cross-database experiments, where the performance drops drastically. In this paper, we analyze the state-of-the-art GC accuracy on three large datasets: MORPH, LFW and GROUPS. We discuss their respective difficulties and bias, concluding that the most challenging and wildest complexity is present in GROUPS. This dataset covers hard conditions such as low resolution imagery and cluttered background. Firstly, we analyze in depth the performance of different descriptors extracted from the face and its local context on this dataset. Selecting the bests and studying their most suitable combination allows us to design a solution that beats any previously published results for GROUPS with the Dago's protocol, reaching an accuracy over 94.2%, reducing the gap with other simpler datasets. The chosen solution based on local descriptors is later evaluated in a cross-database scenario with the three mentioned datasets, and full dataset 5-fold cross validation. The achieved results are compared with a Convolutional Neural Network approach, achieving rather similar marks. Finally, a solution is proposed combining both focuses, exhibiting great complementarity, boosting GC performance to beat previously published results in GC both cross-database, and full in-database evaluations.

Index Terms— Gender classification, HOG, LBP, LSP, LOSIB, information fusion, face local context, cross-database, CNN

1. INTRODUCTION

Gender is a feature easily extracted by humans, and quite useful for human interaction. After all, gender classification (GC) is not yet a solved problem for the computer vision community. Automatic GC is nowadays an active research field
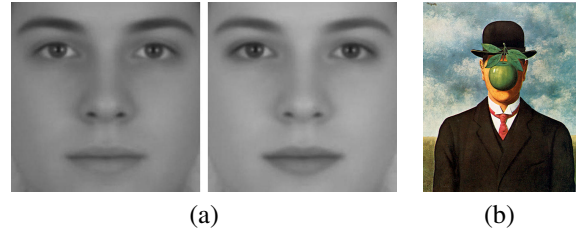
Fig. 1. (a) Images extracted from [1]. Both images are obtained from the same original, but their contrast has been altered, for most human observers the face on the left appears male, while the face on the right appears female. (b) Le Fils de l'homme (René Magritte).

in computer vision, with different application scenarios covering surveillance, demographics or direct marketing among others.

Nowadays, state-of-the-art GC approaches achieve relative high performance based just on visual facial features. However, the exclusive observation of the face might produce perception errors and restrict scenarios of application. In the former, the face pattern may create perception illusions [1], see Figure 1a. The latter occurs in low resolution applications, such as surveillance scenarios, where the facial local context may help to estimate the gender of an individual, see Figure 1b.

1.1. Related work

In this paper, a first objective is to analyze GC results of different datasets to identify those that are closer to real scenarios. As mentioned above, most state-of-the-art GC approaches focus on the facial pattern. This is evidenced by the latest problem surveys [13, 14], and recent results in major journals [2, 9, 10, 15, 16, 17, 18].

As in any classification problem, GC accuracy is estimated on different datasets obtaining the so-called in-database accuracy. However, researchers must not just be interested in getting higher and stable recognition rates for a particularly database, but also improving cross-database

Table 1. Cross-database accuracy rates in the literature evaluating large iage collections: [1] inter-ocular distance > 20, [2] > 20 years old , [3] automatically detected faces of > 20 years old, [4] single face per identity, [5] LFW subset containing 10147 samples.

Reference	Training set	Test set	Accuracy (%)
[2]	FERET	UCN	81.29
[2]	PAL	UCN	74.09
[3]	MORPH	LFW	75.10
[4]	MORPH	LFW	76.64
[5]	GROUPS[1]	LFW	89.77
[6]	GROUPS[1]	LFW	94.48
[7]	GROUPS[2]	LFW	79.53
[8]	GROUPS[3]	LFW	91.62
[9]	GROUPS[3]	LFW[4]	93.35
[10]	4 million faces	LFW	96.86
[11]	CASIA WebFace [12]	LFW[5]	97.10
[3]	MORPH	GROUPS	76.74
[5]	LFW	GROUPS[1]	81.02
[6]	LFW	GROUPS[1]	83.03
[9]	LFW	GROUPS[3]	85.00

Table 2. Recent in-database accuracy results in large datasets. [1] inter ocular distance > 20, [3] 22778 aut. detected faces, [4] 7443 of the total 13233 images, [5] BEFIT protocol, [6] balanced subset with 14244 of the total 55134 images.

Reference	Dataset	Accuracy (%)
[24]	LFW[4]	94.81
[17]	LFW[4]	98.01
[5]	LFW[5]	97.23
[6]	LFW[5]	96.25
[25]	LFW	91.5
[5]	GROUPS[1]	$84.55 - 86.61$
[26]	GROUPS[1]	88.1
[6]	GROUPS[1]	91.59
[27]	GROUPS[1]	92.46
[28]	GROUPS[3]	86.4
[7]	GROUPS[3]	80.5
[29]	GROUPS[3]	90.4
[30]	MORPH	88
[3]	MORPH[6]	97.1

classification rates, i.e. testing with an independent dataset, whose images were captured with different conditions. Cross-database classification is closer to real situations with no dataset bias, that has been proven to provoke optimistic accuracies [7, 19]. Indeed, in real scenarios a gender classifier is trained with a set of images, and later deployed under conditions that may differ from those of the training dataset. Table 1 summarizes recent cross-database classification rates of large databases. We claim that high performance can be obtained particularly in homogeneous, biased and/or reduced datasets, with good image quality, and restricted poses. To illustrate this, different experiments on FERET [20] have recently reported very high GC rates [2, 15, 16]. However, classifiers trained with FERET do not perform robustly with other datasets. The performance drops notoriously [18]. Observing Table 1, with the exception of testing with the biased LFW dataset, the accuracy hardly reaches 80%, evidencing a lack of accuracy in cross-database classification. .

Therefore, part of the community is currently addressing a more realistic or general problem, i.e. GC in the wild. Thus, researchers are now giving more attention to experiments with newer and larger databases that enclose more variability in terms of 1) identity, age and ethnicity, 2) pose and illumination conditions, and 3) image resolution.

Focusing on large and heterogeneous datasets, we highlight recent results reported for the non public UCN [2], and the available MORPH [21], GROUPS [22], and LFW [23] datasets. Observing the in-database rates in Table 2, there is not much room for improvement for LFW and MORPH. We argue that both datasets present some level of simplification that benefits the overall accuracy achieved. In fact, both include multiple samples of the same identity, circumstance that clearly mix gender and identity classification. On the other side, GROUPS offers a less restricted scenario, report-

ing the lowest accuracy, with a large gap compared to other datasets. This evidence has convinced us to focus on this particular dataset, agreeing with the 2015 NIST report conclusion on the topic [14]. We aim at reducing this GC accuracy gap.

Certainly, the need of facial features restricts the context of application, requiring a visible and almost frontal face. From another point of view, different researchers have recently investigated the inclusion of external facial features [28, 31] such as hair, clothing [32, 33, 34], their combination with other cues [35], or even features extracted from the body [36, 37]. The latter claims to be better adapted to real surveillance scenarios where the facial pattern is noisy, not frontal, occluded, or presents low resolution. However, their application is particularly restricted, as no body occlusion may be present.

Indeed, the inclusion of non facial features is consistent with the human vision system that employs external and other features for GC, such as gait, body contours, hair, clothing, voice, etc. [28, 38]. These considerations seem to be of particular interest for degraded, low resolution or noisy images [39]. For those reasons, we will include in our study features extracted from the face, and its local context.

1.2. Contributions

Summarizing, the paper contributions in relation to recent literature are: 1) a study of state-of-the-art GC accuracies in large datasets, analyzing the elements that characterize the current problem challenges, 2) an extensive experimental analysis of a wide collection of descriptors in GROUPS, identifying the best local descriptors for the problem, and an analysis of their robustness against noise; 3) the reduction of the accuracy gap compared to other datasets making use of a score fusion architecture combining features and regions of interest, outperforming previous results, while
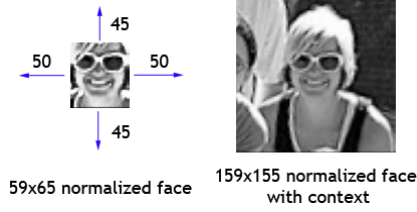
Fig. 2. Normalized face (F) and its corresponding face with context (HS). Each face image is rotated, re-scaled and cropped so that the center of each eye is placed at a fixed location (pixels (16,17) and (42,17) for F). Sample from GROUPS.

confirming that both descriptors and regions provide complementary information for the problem; 7) a further analysis of GC failures for GROUPS; 4) the translation of the approach to cross-database scenarios; including GROUPS, LFW and MORPH. 5) the comparison with a CNN architecture, 6) and the proposal of a solution that combines local descriptors and CNN, boosting significantly previous results;

2. REPRESENTATION AND CLASSIFICATION

Figure 2 illustrates on the left a typical face pattern used for GC. To get this image, a normalization step rotates, scales and crops the image to fix the eyes in specific locations and inter-eye distance (26 pixels), with the resolution of 59×65 pixels. This pattern is referred below as the face (F).

On the right in Figure 2, both the face and its local context are presented. The pattern encloses the hair, shoulders, part of the upper chest, and some background. This is the pattern that we refer hereafter as head and shoulders (HS). The final 159×155 pattern exhibits a large resolution for a real surveillance scenario, and increases the number of features. For that reason and to bring the resolutions closer to those found in real scenarios, in the local descriptors experiments we have considered the HS pattern scaled down to 64×64, 32×32 and 16×16 pixels. Accordingly, in those lower resolution images, the aproximated inter-eye distances are respectively 10, 5 and 3 pixels.

2.1. Representation

Local pixel-wise descriptors have recently received lots of attention with several successful applications to facial analysis [40, 41]. We briefly describe those included in the experimental study.

2.1.1. Histograms of Oriented Gradients (HOG)

HOG [42] is based on the histogram of the gradient orientations in a regular area of the image, called cell. The image is divided into cells, concatenating their respective histograms

to compose the descriptor. In order to reduce the illumination influence, each histogram is normalized using a neighborhood, called block. In their original implementation [42], the cell size is 8×8 pixels and the block is 2×2 cells. The configuration parameters are the number of bins in the histogram, the angle range $0 - 180°$ or $0 - 360°$, the norm used in the normalization stage inside the block, and the overlapping between blocks in the image. In the experiments, we used a 8×8 cells grid, and 9 bins following the implementation by Ludwig et al. [43]. Figure 3 illustrates the gradient orientation in each cell of the image, for a sample at different resolutions.

2.1.2. Local Binary Patterns (LBP)

LBP [44] is a robust and simple but efficient texture descriptor that labels the pixels of an image by thresholding the pixel neighborhood. In texture classification, the LBP code occurrences in an image are described using a histogram.

However, for facial recognition this approximation implies the loss of spatial information. The alternative proposed in [40], divides the face into small regions where the LBP operator is applied and later concatenated into a single histogram. The textures of the facial regions are locally encoded, and the combination of these micro-patterns histograms generates a comprehensive description of the face image.

An extension reduces the dictionary of LBP codes observing the most common ones in texture images. Uniform LBP, LBP^{u2}, contains at most two bitwise transitions from 0 to 1 or vice versa. NILBP [45] is another LBP variant that tries to reduce the LBP local structure oversimplification, computing the difference with the neighbors mean, μ, instead of the central pixel gray value.

A recent redefinition known as Local Salient Patterns (LSP) [46] focuses on the location of the largest differences within the pixel neighborhood. LSP has reported better rates in identity recognition.

Finally, LOSIB [47] is a descriptor enhancer based on LBP that computes local oriented statistical information in the whole image. We have adapted it to face analysis concatenating the histograms obtained from a grid of cells.

The chosen operators are studied below applied on the normalized facial pattern (F) divided into a grid of $n \times n$ cells, in our case, $n = 5$.

2.2. Classification

As the focus of this paper is in the combination of descriptors more than the classifiers themselves, we have used the well known Support Vector Machine (SVM) classifier with RBF kernel [48]. As the reader knows, the SVM classifier obtains the hyperplane that maximizes the class separation to minimize the risk. For non linear separable problems, a previous mapping of the original feature space to a higher dimensional one is carried out by means of kernels.
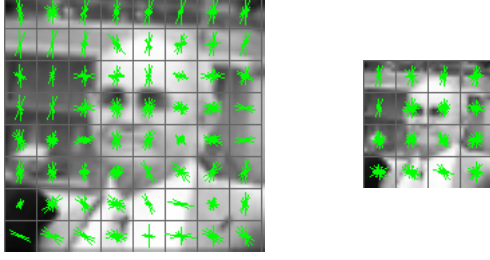
Fig. 3. Relative size of the different HS pattern resolutions considered: respectively 64×64 and 32×32. Their respective HOG grid is depicted.

3. DATABASES

We argued above that an experimental setup with a small or restricted database is not representative for a real world scenario where the gender classifier must face up with thousands of people. To overcome this limitation, we have selected three public databases with a large number of individuals acquired, with the aim at including larger variability. A sample of each dataset in shown in Figure 4, and their respective statistics are presented in Table 3, summarizing their main features as follows:

- **MORPH** [21]. This set contains images of more than $13,000$ identities. We observe however, three undesired features for in the wild scenarios: 1) the number of samples per class is not balanced, 2) the images were acquired indoor in rather similar resolution and illumination conditions (capture bias [49]), and 3) there are multiple samples per individual.

- **Labeled Faces in the Wild** (**LFW**) [23]. The dataset includes images of $5,749$ individuals captured under less controlled conditions. However, 1) it contains several samples per individual, 2) the number of samples per class is not balanced, and 3) the inclusion of public people introduce a selection bias [49].

- **The Images of Groups** (**GROUPS**). This database [22] contains more than $28,000$ labeled faces of lower resolution. According to Table 2, and the recent FRVT report [14], this database is the hardest for GC. Observing Table 3, this is explained due to the lower face resolution (IE) and the larger out of plane rotations ($\bar{\sigma}_{EN}$).

4. EXPERIMENTS

As mentioned before, we firstly focus on GROUPS to analyze the accuracy achieved by a wide collection of descriptors. For this aim, we have followed the experimental protocol defined by Dago et al. [5]. This protocol defines a 5-fold

Fig. 4. Sample images respectively of GROUPS, LFW and MORPH. Their respective original resolutions are 391×293, 249×249 and 200×240 pixels, suggesting a relevant difference in the facial pattern resolution.

Table 3. Databases characteristics: Instances (per class), SNR, inter eye distance mean (IE), and normalized standard deviation of the eye-nose distance ($\bar{\sigma}_{EN}$).

Database	Total (female/male)	SNR	IE	$\bar{\sigma}_{EN}$
GROUPS	28,220 (14,549/13,671)	36.91	25.32 ± 15.4	0.50
LFW	13,232 (2,970/10,252)	36.08	42.16 ± 4.5	0.15
MORPH	55,134 (8,488/46,646)	31.44	92.97 ± 26.4	0.2

cross-validation experimental setup that contains the subset of faces automatically detected with an inter-eye distance larger than 20 pixels[1]. Secondly, we define strategies to improve accuracy and analyze incorrectly classified samples to improve the overall accuracy. Later, aiming at verifying the proposal generalization, we design an in- and cross-database experimental setup that considers the full three selected datasets: GROUPS, LFW and MORPH. Those results are also compared with a CNN designed for GC, to finally conclude the great benefits that the combination of local descriptors and CNN offer to boost GC performance.

4.1. GROUPS Dago's protocol

The results achieved for the Dago's protocol are summarized in Table 4. The experiments have covered the local descriptors collection described above.

The table includes results using both the facial (F), and the head and shoulders (HS) patterns. We remind the reader that F has a resolution of 59×65 pixels, and three different resolutions have been used for HS: 16×16, 32×32 and 64×64, see Section 2 for more details. Observe that the facial resolution contained in HS is up to eight times lower than in F.

For F we employed as descriptors HOG, LBP^{u2}, NILBP, LOSIB, and LSP histograms (respectively F-HOG, F-LBP^{u2}, F-NILBP, F-LOSIB and F-LSP). For HOG, we have selected 8×8 cell histograms with 9 bins. For the rest of descriptors, we have made use of 5×5 histograms, attending our previous experience in [3]. HS is described in terms of HOG

---

and LOSIB features, but considering for the former different resolutions of the pattern.

As baseline, we have included classifiers trained with the first 100 PCA components obtained from the original normalized gray facial images, the histograms obtained from the facial pattern using LBP and HOG, and HS using HOG (F-PCA, F-HOG-PCA, F-LBP-PCA, and $HS_{64\times64}$-HOG-PCA). For classification we present results for SVM+RBF with C (trade-off between margin and error) and gamma values, respectively tuned in the range of $C = [0.25, 8]$ and $gamma = [0.04, 0.15]$.

The best two accuracies for F are provided by: F-HOG (C1), F-LBP (C3), while the best one for HS is reported by $HS_{64\times64}$-HOG (C2). In particular the representation based on F-HOG beats most literature rates of this dataset [5], including our previous results with linear kernels [26]. That descriptor reached 88.23%. It is also remarkable the accuracy achieved using the lowest HS resolution, 75.31%. Observe that Dago et al. [5] made use of a pattern almost twice larger, and the number of features in our previous work was significantly larger. These accuracies were found significantly different ($p = 4 \cdot 10^{-4}$) after carrying out a Kruskal-Wallis test because a previous Jarque-Bera test ($p = 0.5$) rejected the normality of the samples.

The table includes the classification rates per gender, a detail commonly skipped in the literature. In most cases the female accuracy is slightly lower, similarly to the conclusions of the 2015 NIST report [14].

Table 4. GC accuracy (in brackets per class: female/male) achieved using different sets of features (n number of features), resolutions and patterns (F and HS) following the Dago's protocol. Processing time in milliseconds.

Pattern-Descriptor	Accuracy (%)	n	Proc. time
F-PCA	77.91 (77.86/77.95)	100	1.2
F-HOG (**C1**)	**88.23** (88.20/88.25)	576	5.7
F-HOG-PCA	81.10 (81.02/81.38)	100	1
F-LBP^{u2} (**C3**)	86.74 (86.29/87.20)	1475	48.3
F-LBP^{u2}-PCA	80.45 (80.15/80.76)	100	1
F-NILBP	85.31 (85.02/85.59)	1425	48.3
F-LOSIB (**C4**)	86.65 (86.00/87.31)	512	10.4
F-LSP^0	85.58 (84.98/81.17)	1425	39.7
F-LSP^1	85.27 (84.85/85.69)	1425	31.3
F-LSP^2	82.92 (81.94/83.91)	1425	31.3
$HS_{64\times64}$-HOG-PCA	80.80 (80.15/80.76)	100	1
$HS_{64\times64}$-HOG (**C2**)	**85.93** (83.69/88.11)	576	6.2
$HS_{64\times64}$-LOSIB (**C5**)	82.72 (81.41/84.06)	512	10.4
$HS_{32\times32}$-HOG	85.04 (83.13/86.99)	576	11.3
$HS_{16\times16}$-HOG	75.31 (75.08/75.56)	576	18.5

4.2. Robustness against noise

An important element to study, in low resolution scenarios, is the robustness of the proposed approach to the presence of noise. GROUPS contains faces of different resolutions. Of
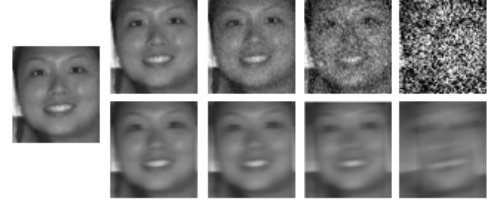

Fig. 5. Original normalized face (59×65), and resulting images after applying Gaussian (first row) or Blurring (second row) noise with different magnitudes: a variance for the gaussian noise up to 0.1, and a linear motion up to 21 pixels for blurring.
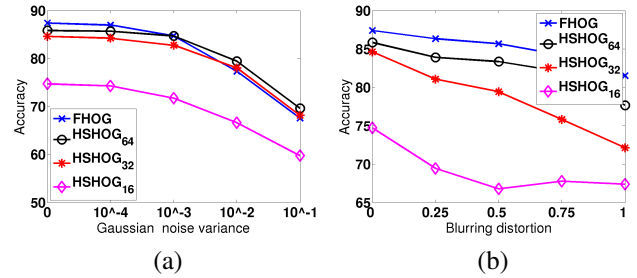

Fig. 6. Accuracies with (a) Gaussian and (b) blur noise.

the total number of samples, a 5% of them present an inter-eye distance lower than 10 pixels, and 41% lower than 20 pixels. Focusing on lower resolution, we include an additional experiment introducing noise to avoid the advantage of getting low resolution images down-sampling higher resolution patterns. Thus, we have noised the images before extracting HOG features.

In Figure 5 we present the original and resulting patterns after applying noise of different nature and magnitude. Figures 6a-b reports the noise influence in GC comparing different pattern resolutions. The accuracy achieved considering only the face pattern (F) is largely affected by the Gaussian noise; reporting lower accuracy than for HS when the noise magnitude increases, see Figure 6a.

4.3. Ensemble of classifiers

Due to the nature of the different descriptors studied, they might provide complementary information. Thus, the combination of all of them in a stacking fashion [50] can improve the overall performance, added to the evidences in the reduction of ambiguous cases occurrences [51].

We explore below score level fusion, as feature level fusion did not report a notorious accuracy improvement, requiring to evaluate a much more complex multi-feature problem. Therefore, we have followed a two stage stacking architecture as illustrated in Figure 7. The first stage obtains the respective output scores of the different single classifiers described in Section 4.1 based on different feature and patterns. The
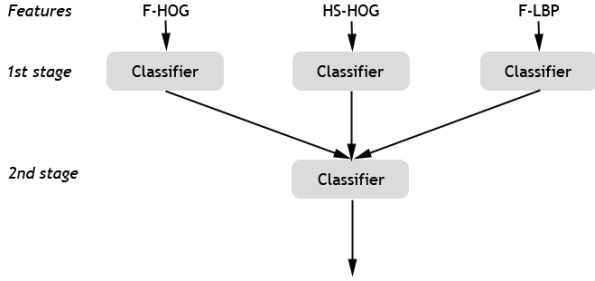
Fig. 7. Illustration of an stacked classifiers architecture, with three classifiers in the first stage whose scores are fed into a second stage "'meta"' classifier.

Fig. 8. ROC curves using the Dago's protocol.

classifier in the second stage considers those scores as inputs.

We have performed an extensive experimental evaluation of possible combinations of those classifiers presented in Table 4. Starting with a fusion of C1 and C2, the best final configuration fuses the information obtained from the following descriptors:

- **C1**. HOG of the facial pattern (F-HOG).

- **C2**. HOG of the head and shoulders pattern ($HS_{64 \times 64}$-HOG).

- **C3**. Concatenated LBP histogram extracted from the facial pattern (F-LBP^{u2}).

- **C4**. Concatenated LOSIB histogram extracted from the facial pattern (F-LOSIB).

- **C5**. Concatenated LOSIB histogram from the head and shoulders pattern ($HS_{64 \times 64}$-LOSIB).

Each first stage classifier is trained using a SVM+RBF. The second stage feeds their respective scores into a new SVM+RBF classifier in charge of taking the final decision. The results achieved for the Dago's protocol are reported in Table 5. In order to confirm the influence of the combination in the results, a Kruskal-Wallis test was carried out and the difference in accuracy was found significant ($p = 4.7 \cdot 10^{-5}$). The ANOVA test was discarded because a previous Jarque-Bera test rejected the normality of the samples ($p = 0.5$).

The results confirm the initial hypothesis of the complementary information contained in the different descriptors. This is evident observing the fusion of features from F, e.g. S1 and S2. There is also a benefit when features are extracted from both F and HS, as evidenced in S3 where LOSIB features are extracted from both patterns, or in S4 and S5. The former fuses LBP and HOG features, the latter integrates also LOSIB features. The accuracy reaches 91.64% (up to 94.28% for adults), beating both the results previously presented in this work, and the literature for this experimental setup. The
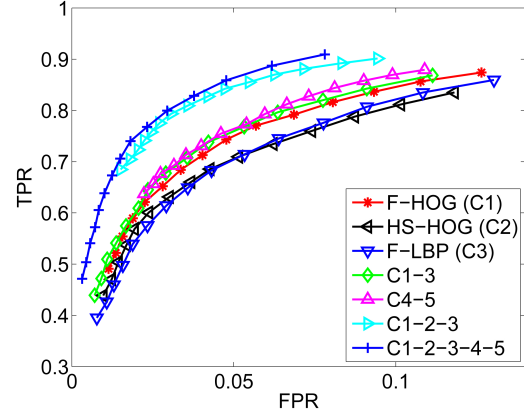
observation of the respective ROC curves, see Figure 8, confirms the best performance of S5.

Table 5. GC accuracy (in brackets female/male) using stacked classifiers for GROUPS.

Classifiers fused	Accuracy (%)
C1-C3 (S1)	88.59 (88.09/89.10)
C1-C3-C4 (S2)	89.18 (88.98 /89.38)
C4-C5 (S3)	88.72 (87.67/89.79)
C1-C2-C3 (S4)	90.44 (90.22/90.66)
C1-C2-C3-C4-C5 (S5)	91.65 (91.05/92.26)
C1-C2-C3-C4-C5 (adults)	**94.28** (94.40/94.16)

The reader may have observed the large improvement when only adults (> 20 years old) are considered in both training and test sets. As recently analyzed, gender discriminant features in children differ from adults [52]. This effect is illustrated in Figure 9, where the GROUPS samples age groups distribution is presented on the left, and the error per gender and age group distribution on the right. Both children and elderly affect negatively the overall recognition accuracy. The former particularly among males, the latter particularly among females. In GROUPS the presence of children is much larger, therefore there is a larger improvement when they are not considered for both test and training, reaching over 94% accuracy.

The influence of elderly is reflected in Figure 10. We present there samples that were incorrectly classified by all the first stage classifiers. Observing the female failures in detail, there is a large presence of elderly ladies, suggesting the inadequate modeling of that particular appearance. Other errors seem to be related with the presence of both glasses, and hats or other elements. The former is affecting the facial features, the latter blocks what may be coped with the local
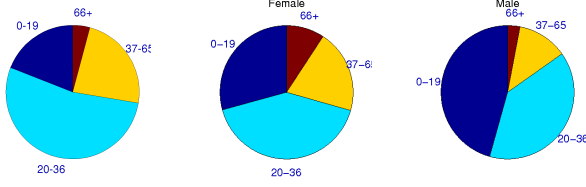
Fig. 9. Ratio (left) of samples per age range in GROUPS, distribution (miggle and right) of errors per age range and class.

Fig. 10. Examples of female samples (HS) with no correct classification.

Fig. 11. Examples of female (left) and male (right) samples correctly classified after adding HS features.

context.

Finally, Figure 11 presents samples (age range $20 - 36$) that were incorrectly classified using the combination S2, i.e. using only the facial features, but correctly classified with the full ensemble of classifiers, i.e. S5, using both facial and local context details. This set of ambiguous facial patterns was better described adding HS features.

4.4. In- and cross-database results in full datasets

In a final experiment, we have tested the best performing classifier in the full selected databases: LFW, GROUPS and MORPH. The in-database (highlighted) and cross-database results are presented in the upper half of Table 6.

Starting with GROUPS, a slight decrease, compared to the Dago's protocol, is observed. Certainly, the inclusion of the whole dataset introduces low resolution and non automatically detected faces, circumstance that adds challenging as-

pects in the experiment. However, the accuracy is closer to 91%, and to 94% if only adults are considered for training and test. Those rates beat any previous results in the dataset, but the best is still to come.

For the other two datasets: MORPH and LFW. We already mentioned that the state-of-the-art literature reports $97 - 98\%$ accuracies. We achieved that accuracy for MORPH, but *only* 95% for LFW. We argue that as stated in those works [17, 24], the authors skipped *faces that are not (near) frontal*, and used higher resolution. The comparison is therefore not completely as fair as it should.
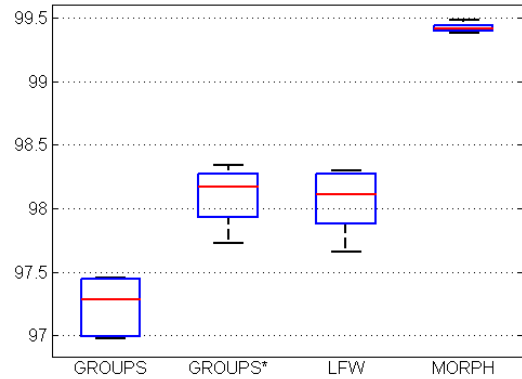

Fig. 12. Boxplot of the accuracy for in-database results (GROUPS* > 20 years old).

Focusing on the most challenging cross-database performance, the observation of Table 6 evidences firstly an improvement compared to previous literature, even if full datasets are evaluated. This suggests that if a more challenging or general problem is carried out, the complementary information contained in both the descriptors, and the regions of interest helps, particularly to tackle large variability, and smaller image resolutions. This achievement is new if compared to the exclusively face centered classifiers.

However, we were not completely satisfied, and analyzed alternatives. CNN [53] have lately achieved relevant results in many Computer Vision problems as image classification [54]. In this sense, some authors have started to evaluate them in GC, with some results reported for LFW and GROUPS [6, 11]. We have adopted the CNN design proposed by [55] with three convolutional layers and two fully connected layers, trained with HS pattern (159×155 pixels), see section 2. The achieved results are summarized in the lower half of Table 6, presenting in most cases slightly better accuracies particularly when testing with GROUPS.

This observation, added to some promising very recent results combining CNN and hand crafted features [6, 56], have guided us to combine local descriptors and CNN, integrating in the proposal another first stage classifier, this time based on

Table 6. GC accuracy (%) with full datasets using hand crafted (left) and CNN (right) classification. The table includes in- (5-folds cross validation) and cross-database results. In-database results are highlighted. [1] > 20 years old

Training set	Test set (hand crafted)		
	GROUPS	LFW	MORPH
GROUPS	90.85	94.10	89.98
GROUPS[1]	93.89	93.94	88.11
LFW	80.22	95	89.16
MORPH	62.04	84.53	98.85

Training set	Test set (CNN)		
	GROUPS	LFW	MORPH
GROUPS	92.90	94.48	87.56
GROUPS[1]	95.82	94.64	90.80
LFW	85.92	96.7	91.84
MORPH	72.32	83.16	98.77

Table 7. GC accuracy (%) combining hand crafted features and CNN.

Training set	Test set		
	GROUPS	LFW	MORPH
GROUPS	97.23	98.00	93.46
GROUPS[1]	98.10	97.95	92.98
LFW	90.14	98.06	93.54
MORPH	67.40	88.70	99.42

CNN. The achieved results are summarized in Table 7. The new results evidence that with the exception of one situation, the one with originally lowest accuracy (MORPH for training, GROUPS for testing), the proposed combination boosted all the accuracies remarkably.

Reviewing first the in-dataset 5-fold cross-validation evaluations, LFW and MORPH reported respectively 98.06% and 99.42%. Thy are indeed new state-of-the-art for both, but certainly similar rates have been achieved for the LFW subset containing almost frontal faces [17, 24]. The accuracy achieved for GROUPS boosted up to 97.23% and to 98.1% when only faces over 19 were used for both training and test. We have no previous referece of any similar reported accuracy neither for the whole dataset, or a subset.

Considering cross-database, training with GROUPS and testing with LFW reported similar numbers to LFW in-dataset results. In fact, previous state-of-the-art for cross-dataset with LFW has already reported 97%, but that was achieved training with 400,000 samples [11] or four millions [10] Here that accuracy is beaten, reaching 98%, and that is done with just a 7% of the training samples used by Antipov et al. When training with MORPH GC rates are lower, 88%, but significantly better that recent reported results that reached 76% [4]. On the other side, GROUPS present larger difficulties, being extremely complex if training with MORPH, just 67%, and easier training with LFW. Again the achieved performance over 90%, is more than 5% better than the latest reported result by

Danisman et al.[9], i.e. 85%

In order to test the significance in performance in the datasets, a previous Jarque-Bera test was carried out to assess the normality of the samples and the results was that the samples are not normally distributed ($p = 0.5$) so a Kruskal-Wallis test was conducted instead of the ANOVA test. As result, it is found a significant difference ($p = 1.1 \cdot 10^{-3}$) in the performance of the datasets, see Figure 12.

4.5. Discussion

In short, we have compared two different focuses to tackle the GC problem. On the one side, we have made use of previous computer vision experience to setup a solution based on local descriptors, that required the almost manual exploration of alternatives, configuration setups, and areas of interest. This focus is now being referred as hand crafted features based, due to the high cost given to the system designer to select the parameters. On the other side, to translate the feature selection and tuning work load to the training process, deep-CNN have also been studied within the same experimental setup.

The evaluation of both solutions reported similar performances, even if a closer look indicates a tiny advantage in terms of accuracy for CNN, compare left and fight accuracies in Table 6.

However, we have proven that the combination at score level of both approaches reaches higher levels of GC accuracy, evidencing better performance both for in- and cross-database evaluations, suggesting the different complementary information for GC provided by descriptors, regions of interest and CNN. For cross-database, previous reported accuracies were typically lower than in-database accuracies. We have however confirmed the claim made by Klare et al. [57] suggesting the importance of demographic variety in the training data. In fact training with GROUPS and testing with LFW reached similar marks than the in-database evaluation for LFW. Under our knowledge, this is the first time that this fact has been made with a relative reduced number of samples.
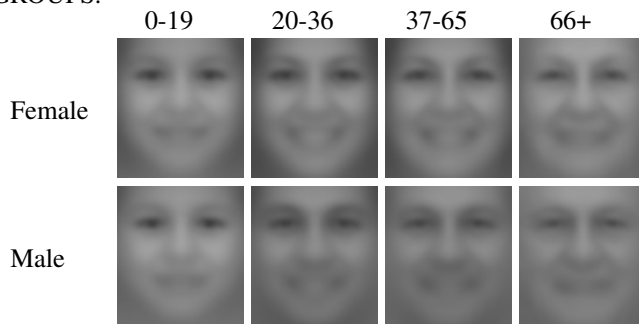
Observing the accuracies achieved for in the wild scenarios, we may wonder whether the problem is solved. Certainly, there is no real evidence that rates over $97 - 98\%$ will be kept after deployment in real world conditions. However, we consider that it is the right time to think about building more challenging in the wild benchmarks to evidence existing difficulties. A very recent work focused on the MORPH dataset [58] presents a joint estimation framework, dealing with the influence of gender and ethnicity on age estimation. Their in-database results achieved high rates for GC ($95 - 98\%$) in MORPH. However, we are concerned with this particular kind of databases that certainly presents a selection bias containing several samples of the same individual. For this reason, we have studied the GC errors in GROUPS. Those age ranges less present produce more classification errors us-

ing local descriptors, while their influence is irrelevant for CNN. However, removing under 20 years old, affects positively both approaches as the population appearance seems to be less spread.

In any case, GROUPS is in fact not ideal, there is a bias in the dataset, that may be illustrated observing the mean face images per age group and gender, see Figure 13. The average faces are smiling, suggesting the particular capture conditions used for the image collection.

Assuming newer and more challenging datasets, we may think about future lines of improvement, that for our proposal may focus on the CNN architecture design, the features and areas of interest analyzed, or even the fusion approach used. In any case, two different GC architecures may be considered as suggested by Klare et al. [57]: a single gender classifier able to handle each demographic group, or a group of classifiers that may be tuned for each demographic group.

Fig. 13. Mean facial patterns per gender and age group in GROUPS.

5. CONCLUSIONS

In this work, we have extensively explored the use of several descriptors and areas of interest for the face based GC problem. The ensemble of different classifiers considering a set of local descriptors and regions of interest in a stacking fashion has proven to reduce almost 20% the previous error rate in the challenging GROUPS dataset following the Dago's protocol.

These results were later confirmed in an experimental evaluation considering in- and cross-database classification in three large datasets: GROUPS, LFW and MORPH. Firstly, the in-database experiment with GROUPS keeps quite similar classification rates, over to 94% in adults. The other two datasets reported similar accuracy rates to the best recent literature.

The experimental evaluation was carried out also for a CNN. The comparison indicates that both approaches perform similarly, with slight advantage for CNN in some experimental scenarios. A further exploration fused both focuses in a score level combination. The accuracies in full in-database cross-validation evaluations for GROUPS, LFW and MOPRH

were respectively boosted up to over 97% (98% in adults), 98% and 99%, reporting also new state-of-the-art accuracies in cross-database GC performance.

6. REFERENCES

[1] Richard. Russell, "A sex difference in facial contrast and its exaggeration by cosmetics," *Perception*, vol. 38, no. 8, pp. 1211–1219, 2009.

[2] Juan Bekios-Calfa, José M. Buenaposada, and Luis Baumela, "Revisiting linear discriminant techniques in gender recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 4, pp. 858–864, April 2011.

[3] Enrique. Ramón-Balmaseda, Javier Lorenzo-Navarro, and Modesto. Castrillón-Santana, "Gender classification in large databases," in *17th Iberoamerican Congress on Pattern Recognition (CIARP)*, 2012, pp. 74–81.

[4] Nesli Erdogmus, Matthias Vanoni, and Sébastien Marcel, "Within- and cross- database evaluations for face gender classification via befit protocols," in *IEEE 16th International Workshop on Multimedia Signal Processing (MMSP)*, 2014, pp. 1–6.

[5] Pablo Dago-Casas, Daniel González-Jiménez, Long Long-Yu, and José Luis Alba-Castro, "Single- and cross- database benchmarks for gender classification under unconstrained settings," in *Proc. First IEEE International Workshop on Benchmarking Facial Image Analysis Technologies*, 2011, pp. 2152–2159.

[6] Jordi Mansanet, Alberto Albiol, and Roberto Paredes, "Local deep neural networks for gender recognition," *Pattern Recognition Letters*, vol. 70, pp. 80–86, 2016.

[7] Juan Bekios-Calfa, José M. Buenaposada, and Luis Baumela, "Robust gender recognition by exploiting facial attributes dependencies," *Pattern Recognition Letters*, vol. 36, pp. 228–234, January 2014.

[8] Taner Danisman, Ioan Marius Bilasco, and Chabane Djeraba, "Cross-database evaluation of normalized raw pixels for gender recognition under unconstrained settings," in *22nd International Conference on Pattern Recognition (ICPR)*, 2014, pp. 3144 – 3149.

[9] Taner Danisman, Ioan Marius Bilasco, and Jean Martinet, "Boosting gender recognition performance with a fuzzy inference system," *Expert Systems with Applications*, vol. 42, pp. 2772–2784, 2015.

[10] Sen Jia and Nello Cristianini, "Learning to classify gender from four million images," *Pattern Recognition Letters*, vol. 58, pp. 35–41, 2015.

[11] Grigory Antipov, Sid-Ahmed Berrania, and Jean-Luc Dugelay, "Minimalistic CNN-based ensemble model for gender prediction from face images," *Pattern Recognition Letters*, vol. 70, pp. 59–65, January 2016.

[12] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z. Li, "Learning face representation from scratch," *arXiv*, vol. 1411.7923, 2014.

[13] Choon-Boon Ng, Yong-Haur Tay, and Bok-Min Goi, "A review of facial gender recognition," *Pattern Analysis and Applications*, vol. 18, pp. 739–755, July 2015.

[14] Mei Ngan and Patrick Grother, "Face recognition vendor test (FRVT) performance of automated gender classification algorithms," Tech. Rep. NIST IR 8052, National Institute of Standars and Technology, April 2015.

[15] Luis A. Alexandre, "Gender recognition: A multiscale decision fusion approach," *Pattern Recognition Letters*, vol. 31, no. 11, pp. 1422–1427, 2010.

[16] Erno Mäkinen and Roope Raisamo, "Evaluation of gender classification methods with automatically detected and aligned faces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 3, pp. 541–547, March 2008.

[17] Juan E. Tapia and Claudio A. Pérez, "Gender classification based on fusion of different spatial scale features selected by mutual information from histogram of lbp, intensity and shape," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 3, pp. 488–499, 2013.

[18] Yasmina Andreu, Pedro García-Sevilla, and Ramón A. Mollineda, "Face gender classification: A statistical study when neutral and distorted faces are combined for training and testing purposes," *Image and Vision Computing*, vol. 32, no. 1, pp. 27–36, January 2014.

[19] Shumeet Baluja and Henry A. Rowley, "Boosting sex identification performance," *International Journal of Computer Vision*, vol. 71, no. 1), pp. 111–119, 2007.

[20] P.Jonathon Phillips, Harry. Wechsler, Jeffery Huang, and Patrick J. Rauss, "The FERET database and evaluation procedure for facerecognition algorithms," *Image and Vision Computing*, vol. 16, no. 5, pp. 295–306, 1998.

[21] Karl Jr Ricanek and Tamirat Tesafaye, "MORPH: A longitudinal image database of normal adult age-progression," in *IEEE 7th International Conference on Automatic Face and Gesture Recognition (FG)*, Southampton, UK, April 2006, pp. 341–345.

[22] Andrew Gallagher and Tsuhan Chen, "Understanding images of groups of people," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 256–263.

[23] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Tech. Rep. 07-49, University of Massachusetts, Amherst, October 2007.

[24] Caifeng Shan, "Learning local binary patterns for gender classification on realworld face images," *Pattern Recognition Letters*, vol. 33, pp. 431–437, 2012.

[25] Yomna Safaa El-Din, Mohamed N. Moustafa, and Hani Mahdi, "Gender classification using mixture of experts from low resolution facial images," in *Multiple Classifier Systems*, vol. 7872 of *Lecture Notes in Computer Science*, pp. 49–60. Springer, 2013.

[26] Modesto Castrillón-Santana, Javier Lorenzo-Navarro, and Enrique Ramón-Balmaseda, "Improving gender classification accuracy in the wild," in *18th Iberoamerican Congress on Pattern Recognition (CIARP)*, 2013, pp. 270–277.

[27] Modesto Castrillón-Santana, Javier Lorenzo-Navarro, and Enrique Ramón-Balmaseda, "On using periocular biometric for gender classification in the wild," *Pattern Recognition Letters*, vol. (in press), 2016.

[28] Neeraj Kumar, Alexander C. Berg, Peter N. Belhumeur, and Shree K. Nayar, "Describable visual attributes for face verification and image search," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, October 2011.

[29] Huizhong Chen, Andrew C. Gallagher, and Bernd Girod, "The hidden sides of names - face modeling with first name attributes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 9, pp. 1860–1873, 2014.

[30] Wen-Sheng Chu, Chun-Rong Huang, and Chu-Song Chen, "Identifying gender from unaligned facial images by set classification," in *International Conference on Pattern Recognition (ICPR)*, Istanbul, Turkey, 2010.

[31] Lubomir Bourdev, Subhransu Maji, and Jitendra Malik, "Describing people: A poselet-based approach to attribute classification," in *International Conference on Computer Vision (ICCV)*, 2011, pp. 1543 – 1550.

[32] Huizhong Chen, Andrew Gallagher, and Bernd Girod, "Describing clothing by semantic attributes," in *European Conference on Computer Vision (ECCV)*, 2012, pp. 609–623.

[33] Bing Li, Xiao-Chen Lian, and Bao-Liang Lu, "Gender classification by combining clothing, hair and facial component classifiers," *Neurocomputing*, vol. 76, no. 1, pp. 18–27, January 2012.

[34] David Freire-Obregón, Modesto Castrillón-Santana, Javier Lorenzo-Navarro, and Enrique Ramón-Balmaseda, "Automatic clothes segmentation for soft biometrics," in *Proceedings of IEEE International Conference on Image Processing (ICIP)*, 2014, pp. 4972–4976.

[35] Abdenour Hadid and Matti Pietikäinen, "Combining appearance and motion for face and gender recognition from videos," *Pattern Recognition*, vol. 42, no. 11, pp. 2818–2827, November 2009.

[36] Matthew Collins, Jianguo Zhang, Paul Miller, and Hongbin Wang, "Full body image feature representations for gender profiling," in *In IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2009, pp. 1235–1242.

[37] Guodong Guo, Guowang Mu, Yun Fu, and Alade Tokuta, "Gender from body: A biologically-inspired approach with manifold learning," in *Ninth Asian Conference on Computer Vision (ACCV)*, 2009, pp. 236–245.

[38] Umar Toseeb, David R. T. Keeble, and Eleanor J. Bryant, "The significance of hair for face recognition," *PLoS ONE*, vol. 7, pp. e34144, 2012.

[39] Pawan Sinha and Tomasso Poggio, "I think I know that face ...," *Nature*, vol. 384, no. 6608, pp. 384–404, 1996.

[40] Timo Ahonen, Abdenour Hadid, and Matti Pietikäinen, "Face description with local binary patterns: Application to face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 12, pp. 2037–204, December 2006.

[41] O Déniz, G Bueno, J Salido, and F De La Torre, "Face recognition using histograms of oriented gradients," *Pattern Recognition Letters*, vol. 32, no. 12, pp. 1598–1603, 2011.

[42] Navneet Dalal and Bill Triggs, "Histograms of oriented gradients for human detection," in *International Conference on Computer Vision & Pattern Recognition (CVPR)*, Cordelia Schmid, Stefano Soatto, and Carlo Tomasi, Eds., June 2005, vol. 2, pp. 886–893.

[43] O. Ludwig, D. Delgado, V. Goncalves, and U. Nunes, "Trainable classifier-fusion schemes: An application to pedestrian detection," in *12th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, October 2009, pp. 1–6.

[44] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[45] L. Liu, P. Fieguth, L. Zhao, Y. Long, and G. Kuang, "Extended local binary patterns for texture classification," *Image and Vision Computing*, vol. 30, no. 2, pp. 86–99, 30.

[46] Zhenhua Chai, Zhenan Sun, Tieniu Tan, and Heydi Mendez-Vazquez, "Local salient patterns - a novel local descriptor for face recognition," in *International Conference on Biometrics (ICB)*, 2013.

[47] O. García-Olalla, E. Alegre, L. Fernández-Roble, and V. González-Castro, "Local oriented statistics information booster (LOSIB) for texture classification," in *International Conference on Pattern Recognition (ICPR)*, 2014.

[48] V. Vapnik, *The nature of statistical learning theory*, Springer, New York, 1995.

[49] Antonio Torralba and Alexei A. Efros, "Unbiased look at dataset bias," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 1521 – 1528.

[50] D. Wolpert, "Stacked generalization," *In Neural Networks*, vol. 5, 1992.

[51] Modesto Castrillón-Santana, Maria De Marsico, Michele Nappi, and Daniel Riccio, "MEG: Multi-Expert Gender classification in a demographics-balanced dataset," in *18th International Conference on Image Analysis and Processing (ICIAP)*, 2015.

[52] R. Satta, J. Galbally, and L. Beslay, "Children gender recognition under unconstrained conditions based on contextual information," in *22nd IEEE International Conference on Pattern Recognition (ICPR)*, Stockholm, Sweden, 2014.

[53] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," in *Proceedings of the IEEE*, 1998, vol. 86, pp. 2278 – 2324.

[54] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, "Imagenet classification with deep convolutional

neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, Eds., pp. 1097–1105. Curran Associates, Inc., 2012.

[55] Gil Levi and Tal Hassner, "Age and gender classification using convolutional neural networks," in *IEEE Workshop on Analysis and Modeling of Faces and Gestures (AMFG), at the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Boston, June 2015, pp. 34–42.

[56] Jos van de Wolfshaar, Mahir F. Karaaba, and Marco A. Wiering, "Deep convolutional neural networks and support vector machines for gender recognition," in *IEEE Symposium Series on Computational Intelligence: Symposium on Computational Intelligence in Biometrics and Identity Management*, 2015.

[57] Brendan Klare and Anil K Jain, "On a taxonomy of facial features," in *Fourth IEEE International Conference on Biometrics: Theory Applications and Systems (BTAS)*, 2010, pp. 1–8.

[58] Guodong Guo and Guowang Mu, "A framework for joint estimation of age, gender and ethnicity on a large database," *Image and Vision Computing*, vol. 32, pp. 761–770, 2014.