Multi-scale score level fusion of local descriptors for gender classification in the wild

M. Castrillón-Santana · J. Lorenzo-Navarro · E. Ramón-Balmaseda

the date of receipt and acceptance should be inserted later

Abstract The 2015 FRVT gender classification (GC) report evidences the problems that current approaches tackle in situations with large variations in pose, illumination, background and facial expression. The report suggests that both commercial and research solutions are hardly able to reach an accuracy over 90% for *The Images of Groups* dataset, a proven scenario exhibiting unrestricted or in the wild conditions. In this paper, we focus on this challenging dataset, stepping forward in GC performance by observing: 1) recent literature results combining multiple local descriptors, and 2) the psychophysics evidences of the greater importance of the ocular and mouth areas to solve this task. We therefore make use of holistic and inner facial patches to extract features, that are later combined via a score level fusion strategy. The achieved results support the main information provided by the ocular and the mouth areas. Indeed, the combination of multiscale extracted features increases the overall accuracy to over 94%, reducing notoriously the classification error if compared with tuned holistic and deep learning approaches.

Keywords Soft biometrics · Gender classification · Local descriptors · Score level fusion · CNN

Work partially funded by the project TIN2015 64395-R from the Spanish Ministry of Economy and Competitiveness, the Institute of Intelligent Systems and Numerical Applications in Engineering (SIANI) and the Computer Science Department at ULPGC. .

SIANI
Universidad de Las Palmas de Gran Canaria
Tel.: +34-928-458743
Fax: +34-928-458711
E-mail: modesto.castrillon@ulpgc.es

1 Introduction

Soft Biometrics have attracted the research community attention, either in scenarios where it is necessary to extract characteristics to describe anonymous individuals, or to improve the performance in identity recognition applications [17,26,38]. We adopt the soft biometrics definition as discrete static and dynamic non unique attributes, that qualitatively separate humans into non-overlapping groups. However, these characteristics lack the distinctiveness and permanence to sufficiently differentiate any two individuals [26]. To illustrate them, we may mention gender, gait, race, mood and apparent age [27,48].

A relevant soft biometric trait is gender. In fact, gender classification (GC) is an active research field with different scenarios of application. Its challenges and relevance have been recently stated in the 2015 Face Recognition Vendor Test (FRVT) related for the first time to this topic [37]. That evaluation points out classification error differences between GC with constrained or controlled datasets, and unconstrained or *in the wild* benchmarks. In the former scenario, standard commercial solutions are able to reach an accuracy around 96.5% with an independent dataset containing roughly one million samples. However in the latter, these solutions are not able to keep a similar performance in smaller datasets, but containing larger range of variation in terms of pose, illumination, etc.

Two theoretically unconstrained datasets were evaluated in the FRVT review: 1) *The Labeled Faces in the Wild* (LFW) [25], and 2) *The Images of Groups* (GROUPS) [20]. On the one side, the best accuracy achieved for LFW was 95.2%, quite close to the numbers reported for constrained datasets. On the other side, the accuracy was significantly lower for GROUPS, hardly reaching 90.4%, suggesting the yet difficult scenario represented by this particular dataset. A similar conclusion has been highlighted by different authors in recent surveys [37,38].

Certainly, GROUPS presents larger variations in terms of pose, background and resolution, see Fig. 1. Additionally, the presence of multiple samples of the same individual is reduced, contrary to LFW, where that circumstance produces an optimistic performance as pointed out by Baluja and Rowley [4].

Moving to the research community results, Table 1 summarizes the most recently reported results for both datasets. Again, the published accuracies suggest that GROUPS is the most complex one. The achieved accuracies are however not comparable to those obtained by commercial systems. Indeed, the research literature results are based on cross-validation, i.e. partitioning the data into complementary sets, repeating for each one the experimental evaluation. Instead, commercial solutions are evaluated using a black-box testing protocol, where there is no knowledge of the system internal structure, i.e. no specific tuning is done for the particular benchmark.

In the search of alternatives to increase GC performance, recently some improvement has been observed making use of strategies that compute features at multiple resolutions of the target area [2,8]. This focus has been combined with the fusion of different descriptors in our recent previous works. In [8]

Fig. 1 Sample images respectively of *The images of Groups* ([20]) and *The Labeled Faces in the Wild* ([25]) datasets. Their respective original resolutions are 391×293 and 249×249 pixels, suggesting a relevant difference in the facial pattern resolution.

the features are extracted from the face and its local context, integrating an additional specialization for the periocular area [11]. Indeed both actions introduce some level of redundancy, but the resulting performance suggests that an adequate design supports an accuracy increase. An additional proven benefit of descriptors fusion for GC is the reduction of the occurrences of ambiguous cases [7]. Considering as ambiguous, those samples reporting a classification score close to the border between the female and male classes.

The aim of this work is to carry out a deeper analysis to evaluate whether the relevant areas of the facial pattern. according to psychophysics. may improve the overall GC accuracy as has already been proved for the periocular area. This work extends [10], firstly almost doubling the number of descriptors evaluated (from 8 to 15). Secondly analyzing the optimal grid resolutions not only for the inner facial areas, i.e. periocular and mouth, but also the whole face and the head and shoulders areas. And finally including in the comparison deep learning based gender classifiers.

The contributions of this work are: 1) independently both the periocular and the mouth area provide an accuracy greater than 80% for GROUPS; 2) finding the optimal grid resolution for each pattern and descriptor; 3) the adequate fusion of features extracted from multi scale facial patterns, reports a significant increase in terms of GC accuracy, reaching over 94% accuracy for GROUPS; and 4) as far as we know, one of the first GC evaluations of Convolutional Neural Networks (CNN) in GROUPS.

2 Related work

In this section, we summarize briefly the approaches whose results are reported in Table 1 for GROUPS and LFW. The reader must observe that the evaluation protocol is not exactly the same in most works. Indeed, it is not

Table 1 GC accuracies in recent research literature for LFW and GROUPS. The whole dataset is used (about 28,000 samples for GROUPS or 13,233 for LFW), with the exception for GROUPS of [1] Dago's protocol containing about 14,000 samples with inter ocular distance > 20, [2] 22,778 automatically detected faces, [3] > 12 years old; and for LFW of [4] 7,443 of the total images, [5] BEFIT protocol, and [6] half dataset.

Reference	Dataset	Evaluation Protocol	Accuracy (%)
[15]	GROUPS	Dago[1]	86.6
[8]			89.8
[9]			91.6
[36]			90.6
[19]			91.6
[11]			92.4
[31]		Detected faces[2]	86.4
[13]			90.4
[6]		Adults[3]	80.5
[23]		Full dataset	87.1
[9]			90.8
[42]	LFW	Subset[4]	94.8
[45]			98.0
[36]		BEFIT[5]	96.2
[15]			94.0
[40]		Half dataset[6]	98.0
[6]		Full dataset	79.5
[41]			94.6
[28]			96.9

common that full datasets are considered, therefore we indicate in such cases the particular subset used in each reference.

A first observation evidences the already mentioned simpler GC problem enclosed in LFW. No matter the experimental protocol used, the achieved accuracies are significantly larger. As mentioned above, the fact that GROUPS is currently the most challenging scenario, is a solid argument to convince us to focus on this particular dataset. Among the different experimental protocols followed by the research community for GROUPS, see Table 1, one of them has been used by different authors as it is reproducible. This experimental protocol was firstly described in Dago et al. [15], and it is adopted below for comparison purposes.

In their work, Dago et al. [15] presented results for LFW and GROUPS evaluating the use of LBP and Gabor features classifying with LDA or SVM. They reported similar accuracies for both features, achieving an accuracy of 86.61% for GROUPS, and 94.01% for LFW. Later, Bekios et al. [5] focused on linear classification combining LDA/PCA features with a Bayesian classifier, integrating gender, age and pose information. They reported an accuracy of 80.5% in GROUPS.

More recently, Shan [42] combined LBP features with a SVM classifier to obtain an accuracy of 94.8% in LFW. Tapia et al. [45] fused different LBP-based features, scales and mutual information measures, reporting for LFW an accuracy of 98%. A similar accuracy has been achieved by Ren and Li [40] and Erdogmus et al. [18]. The former combining two types of local descriptors

(gradient features and Gabor wavelets), and a linear SVM. The latter with LBP -based features. High GC accuracy has also been achieved for LFW making use of an independent training dataset. Jia and Cristianini [28] trained with four million images, achieving 96.9%. More recently Antipov et al. [3] reported 97.1% assembling three Convolutional Neural Networks (CNNs).

Chen and Gallagher [13] built a facial appearance representation based on the 100 most common names in USA. The voting of the top five names is used to assign a gender to a test image. The achieved accuracy for GROUPS reaches 90.4%. Han and Jain [23] employed biologically inspired features (BIF) and a SVM classifier reaching for GROUPS and LFW respectively 87.14% and 95.4%. Slighly higher accuracy, 91.4%, has been achieved by Fazl-Ersi et al. [19] combining LBP, SIFT and color histograms after a feature selection stage.

CNNs [32] have recently reported outstanding results in different Computer Vision problems as image classification [30] and face recognition [43]. An advantage offered by CNN is the reduction of time required for feature selection, as this task is now responsibility of the CNN training process.

Considering these facts, some authors have started to evaluate CNN in GC. We mentioned above the work by Antipov et al. for LFW [3]. However, there is also an interest in combining CNN outputs and local descriptors for GC [47,36]. In particular the proposal by Mansanet et al. [36] weights local patches and Deep Neural Networks (DNN) outputs achieving single dataset GC for LFW and GROUPS respectively of 96.25% and 90.58%.

For comparison purposes we will adopt below the deep CNN design by [33]. In their work the authors made use of three convolutional layers and two fully connected layers for GC and age estimation, creating respectively GenderNet and AgeNet, that were trained and evaluated on the Adience dataset.

3 Proposal

Based on our experience, we keep on exploring the use of multiple descriptors applied to different regions of interest for the GC problem. Our baseline is given by our recent results that compete, as far as we know, the state of the art in facial based GC in the wild [8,9,11], enclosing a comparison with CNN approaches.

Our recent achievement of better accuracies by the fusion of features densely extracted from the face pattern, its local context and some relevant inner facial areas, leads us to revisit the analysis of the human visual system for the GC problem using *bubbles* [22]. In their work, the authors conclude that both the ocular and the mouth areas are significantly discriminant for this particular task to the human system.

Certainly, the use of patches or components is not a novel idea. Indeed, we would like to mention the work by Heisele et al. [24], that made use of two layers of classifiers, being the second the combination of the first layer scores. That approach obtained better results than just using global features. Indeed, our previous study of the integration of the periocular area [11] with

holistic features has evidenced an improvement of the GC performance up to 2 percentage points.

3.1 Fusion strategies

Considering the combination of different descriptors and areas of interest, we may select the fusion strategy that better suits our needs, i.e. fusion either at feature, matching score, or decision level. On one side, feature level fusion will certainly keep more information, but increasing the problem dimensionality will also affect negatively the computational cost. On the completely other side, decision level fusion likely avoids the use of much valuable information. We have therefore adopted a score level fusion (SLF) approach with the aim at achieving the best compromise between speed and performance.

Similarly to Heisele et al., we therefore design a two layered architecture. However, we do not restrict to inner facial regions but also integrate features extracted from the whole facial pattern, and even including its local context. To avoid redundancy and reduce processing cost, a single grid setup is selected for each descriptor and pattern. This selection is done choosing the best performing grid in a previous reduced experiment.

Each first layer expert in the architecture is built using a SVM classifier with RBF kernel [46]. Its output is a score indicating the proximity of the sample to the border between both classes. The second layer fuses those scores feeding a single second layer SVM classifier.

Summarizing, different features are extracted from the selected areas of interest in the first layer, to apply a score level fusion (SLF) in the second step. SLF is adopted considering the benefits involved in managing smaller feature vectors, and the possibility of parallel computation.

3.2 Regions of interest

The different regions of interest considered are illustrated in Fig. 2: head and shoulders (HS), face (F), periocular (P), and mouth (M) areas. All those patterns are automatically cropped from the original head and shoulders pattern (with a dimension of 155×159 pixels where the inter-eye distance has been normalized to 26 pixels). A previous image normalization process is guided by the eyes, that encloses rotation, scaling and translation to fix the eye locations.

3.3 Descriptors

As features, we do not define any new descriptor, but evaluate a collection of local descriptors. Local descriptors are currently being extensively applied for facial analysis, making use of a histogram representation to reduce the feature vector dimension. However, typically a grid of homogeneous cells is defined

Fig. 2 GROUPS sample presenting the regions of interest considered below, respectively and starting from the left: head and shoulders (HS) (64×64 pixels), face (F) (59×65 pixels), periocular (P) (49×19 pixels), and mouth (M) (37×31 pixels).

to avoid the loss of spatial information produced by a single based histogram representation [1].

A grid resolution is defined by its number of horizontal and vertical cells, respectively cx and cy. Therefore, the pattern is divided into a total of $cx \times cy$ cells. For a given descriptor, a histogram is computed in each cell, h_i, where the bins indicate the number of occurrences of the different codes. The final feature vector, $\mathbf{x}$, is composed concatenating the respective $cx \times cy$ histograms, i.e. $\mathbf{x} = \{h_1, h_2, ..., h_{cx \times cy}\}$.

Summarizing, we evaluate each pattern for each descriptor with a particular range of grid configuration. For all the patterns we have covered the range $cx \in [1, 8]$ and $cy \in [1, 8]$. For the periocular pattern, P, due to its narrower height, we have restricted the study to the range $cx \in [1, 8]$ and $cy \in [1, 6]$. That makes respectively a total of 64 and 48 variants per descriptor. The best grid resolution for each particular local descriptor is later used for fusion strategies. As descriptors we have considered the well known HOG and LBP, and some alternatives, including different LBP variants. We include a brief description, with some additional details for those less commonly used for GC:

- Histogram of Oriented Gradients (HOG) [16]. Based on the gradient orientations in each image cell. As mentioned above, and similarly to the whole collection of descriptors, an image is represented by the concatenation of the respective cell histograms. Illumination normalization is applied for each cell histogram, considering its neighborhood, known as block. We have adopted the implementation by [35] that considers blocks of 2×2 cells, and 9 bin histograms.
- Local Binary Patterns (LBP) and uniform Local Binary Patterns (LBP^{u2}) [1]. LBP is a robust texture descriptor that encodes a pixel attending to whether its gray value is greater or not each of its neighbors, composing a binary code. Its generalized definition for an arbitrary circular neighborhoods of radius R with P neighbors is:

$$LBP_{P,R} = \sum_{p=0}^{P-1} s(g_p - g_c)2^p \tag{1}$$

where, g_c is the gray level of the central pixel and g_p (with $p = 0, 1, \cdots, P-1$) are the values of its P neighbors. The function $s(x)$ is defined as:

$$s(x) = \begin{cases} 1, x \geq 0 \\ 0, x < 0 \end{cases} \tag{2}$$

LBP^{u2} reduces the codes dictionary observing those codes more common in texture images.

- Local Gradient Patterns (LGP) [29]. As an alternative to the LBP definition, LGP descriptor makes use of the neighborhood gradient to encode the image pixel. The gradient is computed as the absolute value of intensity difference between the central pixel and each of its neighborhood pixels. Similarly to LBP, its definition for an arbitrary circular neighborhoods of radius R with P neighbors:

$$LGP_{P,R} = \sum_{p=0}^{P-1} s(g_p - \bar{g})2^p \tag{3}$$

where, $\bar{g}$ is the threshold value, commonly the gradient mean, and g_p ($p = 0, 1, \cdots, P-1$) are the gradient values of the P neighbors of the central pixel as function $s(x)$ is defined in Eq. 3.3.

- Local Ternary Patterns (LTP) [44]. Compared to LBP, LTP considers three possible relations instead of two obtaining a ternary code. In fact, an intensity range, $\pm t$, is defined around the gray level g_c of the central pixel Gray levels within that range are quantized to 0, while those below $g_c - t$ to -1, and those above $g_c + t$ to 1. In its definition, the function $s(x)$ is:

$$s(x) = \begin{cases} 1, x \geq g_c + t \\ 0, |x - g_c| < t \\ -1, x \leq g_c - t \end{cases} \tag{4}$$

Theoretically, LTP is more robust to noise. The resulting ternary may be split into its positive and negative parts (below LTP_{high} and LTP_{low}) that can be used as separate descriptors.

- Local Salient Patterns (LSP) [12]. This LBP alternative focuses on the largest differences computed within the pixel neighborhood. This is done to remove noise influence. The basic idea is to obtain the pair of neighbor indexes ($p_{diffmax}, p_{diffmin}$) indicating respectively the maximum and minimum differences with the central value. Considering a 3×3 neighborhood, five different variants are evaluated.
  - LSP_0 refers to the gray level difference of each pixel with respect to the window central pixel.

  - LSP_1 computes the difference of each pixel with respect to the following one in the circular neighborhood.
  - LSP_2 computes for each p_i in the neighborhood, the (circular) value $p_i + p_{i+2} - 2p_{i+1}$
  - LSP_{01} merges the results of LSP_0 and LSP_1.
  - LSP_{012} concatenates the results of all three LSP_n.
- Weber Local Descriptor (WLD) [14]. WLD also encodes differences of pixel intensity within a local neighborhood comprising differential excitation and orientation. In this sense, WLD is based on Weber's Law stating that human perception of a pattern depends both on the change of a stimulus and also on its original intensity.
- Local Phase Quantization (LPQ) [39]. Insensitive to centrally symmetric blur, it is based on the blur invariance property of the Fourier phase spectrum. In this sense, this descriptor is computed using the short-term Fourier transform (STFT) within the neighborhood.
- Intensity based Local Binary Patterns (NILBP) [34]. This LBP variant focuses on the reduction of the LBP oversimplification of local structure. This is done computing the difference of each neighborhood pixel with the neighborhood mean, μ, instead of the central pixel gray value.
- Local Oriented Statistics Information Booster (LOSIB) [21]. Texture enhancer based on LBP, that computes the local oriented statistical information in the whole cell. This is done first computing the intensity differences in the 3×3 neighborhood as follows:

$$d_k(x_c, y_c) = |g_k - g_c| \tag{5}$$

  being $k = 0, 1, ..., P-1$ and P the pixel neighbors. LOSIB computes the mean of all differences along the p orientations for the $m \times n$ cell pixels:

$$\nu_k = \frac{\sum_{x_c=1}^{m} \sum_{y_c=1}^{n} d_k(x_c, y_c)}{m \cdot n} \tag{6}$$

  Each cell is described in terms of P mean values, i.e. $\{\nu_0, \nu_2, ..., \nu_{p-1}\}$.

4 Experiments and results

We adopt the Dago's experimental protocol, that defines a 5-fold cross validation for GROUPS. The protocol reduces the original GROUPS dataset, including only faces with an inter-eye distance larger than 20 pixels, making a total of about $14,000$ samples. We remind the reader that those faces are then normalized to inter-eye distance of 26 pixels before extracting the studied patterns, see Fig. 2.

Firstly, the study focuses on each pattern (F, HS, P and M) individually, analyzing the optimal descriptor grid setup for their resolution. Observe that

only the HS pattern is down-sampled to 64×64 pixels to make it more manageable. Secondly, fusion strategies are evaluated, initially considering a single pattern, to later combine multiple descriptors and patterns. We summarize the experimental steps as follows:

1. Explore grid resolutions for each particular pattern and feature in the first Dago's fold.
2. Evaluate the fusion of descriptors for a given pattern
3. Evaluate the fusion of descriptors extracted from F and P.
4. Evaluate the fusion of descriptors extracted from F and M.
5. Evaluate the fusion of descriptors extracted from F, P and M.
6. Evaluate the fusion of descriptors extracted from F and HS
7. Evaluate the fusion of descriptors extracted from F, HS and P.
8. Evaluate the fusion of descriptors extracted from F, HS and M.
9. Evaluate the fusion of descriptors extracted from F, HS, P and M.
10. Evaluate GC with available CNN based on the HS pattern.

4.1 Single descriptor and pattern

We present in first place the results achieved making use of features extracted only from each individual pattern. Table 2 summarizes the accuracies achieved for the first Dago's fold using the whole collection of descriptors and variants. The table includes only the best grid configuration accuracy for each descriptor and pattern.

As expected both F and HS reported the highest numbers, respectively over 87% and 85%. For F that accuracy was achieved using HOG, while a group of other descriptors reported over 86%. Considering HS, there are indeed at least 7 descriptors with an accuracy larger than 85%. The leader is LBP^{u2} but also HOG, WLD, NILBP and some LSP variants are rather close.

Observing the other patterns, i.e. P and M, it is evident that their accuracies are significantly worse than those reported by recent face based GC systems.

The addition of a larger collection of local descriptors introduces some slight differences in best accuracies with precedent results. Compared to our previous work [10], where P and M were analyzed, the P pattern gets the best accuracy using LSP_{012} with 81.77%, slightly better than using HOG. For M pattern HOG keeps being the leader, followed by WLD. The newer results for F and HS did not bring any improvement in the overall best accuracy. In terms of accuracy, the patterns may be sorted in descending order as F, HS, P and M.

Finally, we would like to make the reader observe a single descriptor such as HOG. The optimal selected grid evidences that for each pattern the cells are covering different extension over the facial pattern, suggesting indeed a multi scale analysis. Certainly, the optimal grids selected for P and M, cover some areas with larger detail, that is quite unpractical for the whole pattern, as the feature vector dimension would increase significantly.

Table 2 Summary per pattern using a single descriptor. For each pattern and descriptor, the number of features (#), mean processing time per image (milliseconds for a Matlab implementation) and grid resolution are presented added to the accuracy. Achieved accuracies are highlighted, the darker the color, the worse the accuracy.

Descriptors	F			HS		
	# (grid)	t	Acc.	# (grid)	t	Acc.
HOG	576 (8×8)	9	**87.48**	504 (7×8)	9	85.62
LBP^{u2}	1770 (6×5)	71	86.78	1770 (6×5)	75	**85.75**
LBP	3072 (3×4)	96	82.90	3840 (3×5)	138	82.07
LGP	14436 (7×8)	254	85.31	16384 (8×8)	316	83.19
LPQ	1024 (1×4)	21	85.14	3072 (4×3)	56	84.14
WLD	6400 (5×5)	356	86.60	4608 (3×6)	316	85.40
LOSIB	448 (8×7)	7	83.00	384 (6×8)	6	82.64
NILBP	2478 (7×6)	84	86.20	2832 (8×6)	96	85.66
LSP_0	1710 (6×5)	107	85.70	1995 (7×5)	120	85.45
LSP_1	2394 (6×7)	117	85.98	1710 (6×5)	93	85.13
LSP_2	2280 (5×8)	111	83.64	2052 (6×6)	104	84.47
LSP_{01}	2736 (6×4)	149	86.58	2280 (4×5)	130	85.65
LSP_{012}	2052 (3×4)	122	86.28	3078 (6×3)	162	84.64
LTP_{high}	3072 (3×4)	103	86.06	3072 (4×3)	118	84.16
LTP_{low}	2048 (2×4)	69	85.34	3072 (4×3)	119	83.52

	P			M		
	# (grid)	t	Acc.	# (grid)	t	Acc.
HOG	576 (7×6)	14	81.61	378 (8×8)	8	**80.50**
LBP^{u2}	1475 (8×3)	49	79.60	1416 (5×5)	41	76.76
LBP	5120 (6×3)	115	78.30	4608 (4×5)	93	75.06
LGP	10752 (6×6)	179	77.09	9216 (7×6)	138	76.66
LPQ	1024 (2×2)	23	77.12	1024 (2×2)	17	77.47
WLD	5120 (6×3)	177	81.03	4608 (4×5)	148	78.42
LOSIB	336 (7×6)	7	76.09	336 (7×6)	7	72.79
NILBP	1770 (7×3)	51	78.67	1239 (6×5)	33	77.10
LSP_0	1425 (8×3)	49	78.64	1368 (5×5)	40	76.59
LSP_1	1425 (7×3)	48	78.77	1197 (5×5)	36	76.92
LSP_2	1710 (7×3)	62	76.72	1197 (6×5)	35	73.05
LSP_{01}	2280 (6×2)	79	80.49	1368 (5×4)	45	78.30
LSP_{012}	2565 (8×2)	93	**81.77**	2736 (5×3)	85	78.47
LTP_{high}	3072 (6×2)	75	80.25	3072 (6×2)	6	77.90
LTP_{low}	7168 (7×2)	159	80.36	3584 (4×7)	77	77.90

Table 3 Summary per pattern fusing descriptors of a single pattern.

Pattern	Acc.	Descriptors
F	**89.22**	HOG + LGP + LPQ
HS	88.68	LBP^{u2} + WLD + LSP_{01}
P	83.41	WLD + LSP_{012} + LTP_{high}
M	81.58	HOG + LSP_0 + LSP_1

4.2 Multiple descriptors and single pattern

Considering SLF, we first evaluate the combination of multiple descriptors for a single pattern. However, we have limited the number of descriptors combined to a maximum of three to reduce the search space, as for each pattern there are 2^{15} possible combinations. Table 3 presents the best scores achieved combining multiple descriptors per pattern after computing the mean for the 5-folds, varying the cost and gamma parameters respectively within the intervals $C = [0.5, 5]$ and $gamma = [0.04, 0.15]$. Again F and HS are clearly reporting the highest accuracies, rather close to those present in recent literature, but this time the difference among them is reduced, even when the facial pattern presents a remarkable lower resolution within the HS pattern. They respectively have increased roughly 1.74% (from 87.48% to 89.22%) and 2.93% (from 85.75% to 88.68%) percentage points. P and M also increased their respective accuracies combining descriptors, but they are still far from the scores achieved by F and HS.

Observe that the best combinations do not necessarily make use of the same descriptors, and those combined are not necessarily the individually best ones. These results suggest the complementary information contained in some descriptors for this problem. However, some descriptors appear more often (HOG, WLD and LSP variants).

4.3 Multiple descriptors and patterns

The last SLF experiment evaluates the combination of multiple descriptors and patterns. Certainly, the analysis of all possible combinations is far from being tractable, once more we therefore have limited the space search. Observing their respective individual rates and presence in the previous section best combinations, for face (F) only HOG and LBP^{u2} are considered; for head and shoulders (HS) HOG, LBP^{u2} and WLD; for periocular (P) HOG, LBP^{u2}, LGP, LPQ, WLD, LSP_{012}, LTP_{low} and LTP_{high}; and for mouth (M) HOG, WLD, LSP_0 and LSP_1. Additionally, no more than three descriptors per pattern are combined.

Table 4 summarizes the results. Similarly to the previous section, the reported results correspond to the 5-folds mean highest accuracy achieved, varying the cost and gamma parameters respectively within the intervals $C = [0.5, 5]$ and $gamma = [0.04, 0.15]$. These results are compared with a holistic approach with similar accuracy to the recent literature (F+HS), and our previous results integrating ocular features.

The combination of all patterns reaches an accuracy over 94%. Compared to Table 1 where no major attention was given to the periocular or mouth areas, the accuracy is increased more than 2.4 percentage points, and the error reduction is almost 30%. This combination integrates in the first layer 10 descriptors. Among them the slowest is HS-WLD requiring 316 msec. per image. Even when each descriptors may be computed in parallel, when processing

Table 4 Mean accuracies (in brackets female/male) for Dago's protocol with SLF based on the face (F), head and shoulders (HS), periocular (P) and mouth (M) areas. The first part presents results for a single pattern, the second combining patterns and descriptors, and the last part includes recent literature and CNN results for comparison purposes. Each result is associated with the pattern and features fused.

Pattern(s)	Approach	Descriptors	Acc.
F	Single	F-HOG	87.48 (87.45/87.50)
	SLF	F-HOG + F-LGP F-LPQ	89.22 (89.22/89.21)
HS	Single	HS-LBP^{u2}	85.75 (84.53/87.01)
	SLF	HS-LBP^{u2} + HS-WLD HS-LSP$_{01}$	88.68 (88.79/88.57
P	Single	P-LSP$_{012}$	81.77 (81.08/82.47)
	SLF	P-WLD + P-LSP$_{012}$ P-LTP$_{high}$	83.41 (82.68/81.53)
M	Single	M-HOG	80.55 (80.73/80.37)
	SLF	M-HOG + M-LSP$_0$ M-LSP$_1$	81.58 (78.99/83.15)
F+P	SLF	F-HOG + F-LBP^{u2} P-HOG + P-LOSIB P-LPQ	91.22 (91.89/90.53)
F+M	SLF	F-HOG + F-LBP^{u2} M-HOG + M-LSP$_1$	90.46 (90.59/90.33)
F+P+M	SLF	F-HOG + F-LBP^{u2} P-HOG + P-LPQ P-LBP^{u2} M-HOG + M-LSP$_1$	92.22 (92.46/91.98)
F+HS	SLF	F-HOG + F-LBP^{u2} F-LGP + HS-HOG HS-LBP^{u2}+ HS-WLD	91.12 (91.11/91.13)
F+HS+P	SLF	F-HOG + F-LBP^{u2} HS-LBP^{u2} + HS-WLD P-HOG + P-LOSIB	93.54 (93.78/93.29)
F+HS+M	SLF	F-HOG + F-LBP^{u2} HS-LBP^{u2} HS-WLD + M-HOG M-LBP^{u2} + M-WLD	93.40 (93.53/93.26)
F+HS+P+M	SLF	F-HOG + F-LBP^{u2} HS-HOG + HS-LBP^{u2} HS-WLD+ P-HOG P-LPQ + P-LOSIB M-HOG + M-LSP$_1$	**94.04** (94.30/93.78)
F+HS	SLF [9]	F-HOG + F-LBP^{u2} HS-HOG F-LOSIB	91.6
F+HS+P+M	SLF [10]	F-HOG + F-LBP^{u2} HS-HOG + P-HOG P-LBP^{u2} M-HOG + M-WLD	93.22
F+HS+P	SLF [11]	F-HOG + F-LBP^{u2} HS-HOG + P-HOG P-LBP^{u2} + P-LOSIB	92.46
HS	ImageNet [30]	fc6	87.98
HS	GenderNet [33]	-	80.42
HS (159 × 155)	GenderNet [33] trained with GROUPS	- -	92.89
Face (60 × 60)	Weighted combination [36]	Local-DNN	90.58
Face	Pool of features [19]	SIFT, LBP, Color histogram	91.59

cost is a restriction slow descriptors may be avoided during the search process of the best combination.

Observing the accuracies and a relevant subset of the corresponding ROC curves in Fig. 3 the proposal reports the best AUC. Indeed, the fusion of F and HS, respectively with P or M (but not both) reports quite similar results. Both are indeed better that our previous analysis with a reduced collection of local descriptors combining both patterns [10].

A final comment may be given observing the reported accuracy per class in Table 4. The SLF of multiple descriptors and patterns, reports slightly better accuracy for the female class in most cases compared to single pattern results.

4.4 Convolutional Neural Networks

As already mentioned, CNNs [32] have achieved very high performance score in many Computer Vision problems as image classification [30]. According to our knowledge, they have rarely been applied to GROUPS for GC. We have evaluated both a general (Imagenet) and a specifically trained (GenderNet) CNN using our full HS pattern (159×155 pixels).

Due to the CNN standard structure, stacked convolutional layers followed by one or more fully connected layer that acts as MLP classifier, they can be considered as feature extractors with the resulting features as the fully connected layer output. In this sense, the output of the fc6 layer of the pretrained Imagenet [30] has been considered as a vector of 4096 features to feed a SVM classifier (RBF kernel, C= 5, gamma=0.06). The results of this approach are included in the bottom part of Table 4. The achieved accuracy of 87.98% reaches similar numbers to the best single descriptors results obtained for F and HS, beating any solution exclusively based on P or M.

On the other hand, GenderNet is a CNN proposed by [33] with 3 convolutional layers and 2 fully connected layers trained with the dataset Adience benchmark where the authors obtained the best results till then. This specifically trained CNN has been used as an end-to-end classifier. The results are shown in Table 4. The accuracy, 80.42%, is the lowest of the compared approaches, evidencing the difficulties of GROUPS. When GenderNet is trained with GROUPS, the achieved accuracy increases up to 92.89%, beating the results by Mansanet et al. [36] but still lower than our proposal including multiple descriptors and regions of interest.

5 Conclusions

This paper explores the benefits of combining holistic features with features extracted from specific inner facial regions in the context of GC. For that purpose, we have extracted features from the whole facial pattern, its local context, and two inner facial regions of great relevance for the human visual system: the ocular and mouth areas. In fact, this approach consists in extracting features at different resolutions from those specific facial areas.

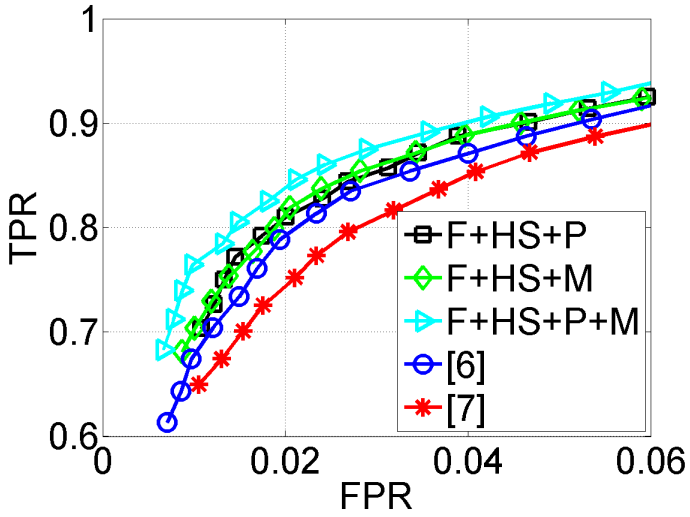

Fig. 3 ROC curves using Dago's protocol. Comparison of state-of-the-art classification based on F and HS, with the proposed fusion alternatives considering HS and F features respectively with P, M and both.

In relation to our previous work, we have extended the collection of local descriptors studied. The achieved results report a significant improvement in terms of accuracy and classification error reduction, when descriptors from multiple areas are fused by means of score level strategies for this particular problem. This fusion strategy is well suited to reduce the feature vector dimension and allow parallel computation. This approach is also compared with CNN implementations, ImageNet and GenderNet, suggesting that GC solutions based on hand crafted features may still compete with deep-CNN.

Summarizing, GC error is remarkably reduced if added to facial information, features are specifically extracted from the periocular and mouth areas. Both CNN and hand crafted features based approaches based exclusively on holistic patterns report accuracies up to 92%. However, the multiscale proposal fusion of descriptors and areas of interest reach an accuracy over 94% for GROUPS. These results reduce significantly the performance gap compared to GC on controlled datasets.

References

1. Ahonen, T., Hadid, A., Pietikäinen, M.: Face description with local binary patterns: Application to face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **28**(12), 2037–204 (2006)
2. Alexandre, L.A.: Gender recognition: A multiscale decision fusion approach. Pattern Recognition Letters **31**(11), 1422–1427 (2010)
3. Antipov, G., Berrania, S.A., Dugelay, J.L.: Minimalistic CNN-based ensemble model for gender prediction from face images. Pattern Recognition Letters **70**, 59–65 (2016)

4. Baluja, S., Rowley, H.A.: Boosting sex identification performance. International Journal of Computer Vision **71**(1)), 111–119 (2007)
5. Bekios-Calfa, J., Buenaposada, J.M., Baumela, L.: Revisiting linear discriminant techniques in gender recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **33**(4), 858–864 (2011)
6. Bekios-Calfa, J., Buenaposada, J.M., Baumela, L.: Robust gender recognition by exploiting facial attributes dependencies. Pattern Recognition Letters **36**, 228–234 (2014)
7. Castrillón-Santana, M., De Marsico, M., Nappi, M., Riccio, D.: MEG: Multi-Expert Gender classification in a demographics-balanced dataset. In: 18th International Conference on Image Analysis and Processing (ICIAP) (2015)
8. Castrillón-Santana, M., Lorenzo-Navarro, J., Ramón-Balmaseda, E.: Improving gender classification accuracy in the wild. In: 18th Iberoamerican Congress on Pattern Recognition (CIARP), pp. 270–277 (2013)
9. Castrillón-Santana, M., Lorenzo-Navarro, J., Ramón-Balmaseda, E.: Descriptors and regions of interest fusion for gender classification in the wild. ArXiv e-prints (2015)
10. Castrillón-Santana, M., Lorenzo-Navarro, J., Ramón-Balmaseda, E.: Fusion of holistic and part based features for gender classification in the wild. In: New Trends in Image Analysis and Processing–ICIAP 2015 Workshops, pp. 43–50. Springer International Publishing (2015)
11. Castrillón-Santana, M., Lorenzo-Navarro, J., Ramón-Balmaseda, E.: On using periocular biometric for gender classification in the wild. Pattern Recognition Letters **(in press)** (2016). DOI http://dx.doi.org/10.1016/j.patrec.2015.09.014
12. Chai, Z., Sun, Z., Tan, T., Mendez-Vazquez, H.: Local salient patterns - a novel local descriptor for face recognition. In: International Conference on Biometrics (ICB) (2013)
13. Chen, H., Gallagher, A.C., Girod, B.: The hidden sides of names - face modeling with first name attributes. IEEE Transactions on Pattern Analysis and Machine Intelligence **36**(9), 1860–1873 (2014)
14. Chen, J., Shan, S., He, C., Zhao, G., Pietikainen, M., Chen, X., Gao, W.: Wld: A robust local image descriptor. Pattern Analysis and Machine Intelligence, IEEE Transactions on **32**(9), 1705–1720 (2010). DOI 10.1109/TPAMI.2009.155
15. Dago-Casas, P., González-Jiménez, D., Long-Yu, L., Alba-Castro, J.L.: Single- and cross- database benchmarks for gender classification under unconstrained settings. In: Proc. First IEEE International Workshop on Benchmarking Facial Image Analysis Technologies, pp. 2152–2159 (2011)
16. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: C. Schmid, S. Soatto, C. Tomasi (eds.) International Conference on Computer Vision & Pattern Recognition (CVPR), vol. 2, pp. 886–893 (2005)
17. Dantcheva, A., Elia, P., Ross, A.: What else does your biometrics data reveal? A survey on soft biometrics. IEEE Transactions on Information Forensics And Security **11**, 441–467 (2016)
18. Erdogmus, N., Vanoni, M., Marcel, S.: Within- and cross- database evaluations for face gender classification via befit protocols. In: IEEE 16th International Workshop on Multimedia Signal Processing (MMSP), pp. 1–6 (2014)
19. Fazl-Ersi, E., Mousa-Pasandi, M.E., Laganiere, R., Awad M., .: Age and gender recognition using informative features of various types. In: International Conference on Image Processing (2014)
20. Gallagher, A., Chen, T.: Understanding images of groups of people. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), pp. 256–263 (2009)
21. García-Olalla, O., Alegre, E., Fernández-Roble, L., González-Castro, V.: Local oriented statistics information booster (LOSIB) for texture classification. In: International Conference on Pattern Recognition (ICPR) (2014)
22. Gosselin, F., Schyns, P.G.: Bubbles: a technique to reveal the use of information in recognition tasks. Vision Research **41**(17), 2261–2271 (2001)
23. Han, H., Jain, A.K.: Age, gender and race estimation from unconstrained face images. Tech. Rep. MSU-CSE-14-5, Michigan State University (2014)
24. Heisele, B., Serre, T., Poggio, T.: A component-based framework for face detection and identification. International Journal of Computer Vision Research **74**(2) (2007)

25. Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Tech. Rep. 07-49, University of Massachusetts, Amherst (2007)
26. Jain, A.K., Dass, S.C., Nandakumar., K.: Soft biometric traits for personal recognition systems. In: International Conference on Biometric Authentication, pp. 731–738 (2004)
27. Jain, A.K., Kumar, A.: Secong Generation Biometrics, chap. Biometrics of Next Generation: An Overview, pp. 49–79. Springer (2012)
28. Jia, S., Cristianini, N.: Learning to classify gender from four million images. Pattern Recognition Letters **58**, 35–41 (2015)
29. Jun, B., Kim., D.: Robust face detection using local gradient patterns and evidence accumulation. Pattern Recognition **45**(9), 3304–3316 (2012)
30. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: F. Pereira, C. Burges, L. Bottou, K. Weinberger (eds.) Advances in Neural Information Processing Systems 25, pp. 1097–1105. Curran Associates, Inc. (2012)
31. Kumar, N., Berg, A.C., Belhumeur, P.N., Nayar, S.K.: Describable visual attributes for face verification and image search. IEEE Transactions on Pattern Analysis and Machine Intelligence pp. 1962–1977 (2011)
32. Lecun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. In: Proceedings of the IEEE, vol. 86, pp. 2278 – 2324 (1998)
33. Levi, G., Hassner, T.: 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), pp. 34–42. Boston (2015)
34. Liu, L., Fieguth, P., Zhao, L., Long, Y., Kuang, G.: Extended local binary patterns for texture classification. Image and Vision Computing **30**(2), 86–99 (30)
35. Ludwig, O., Delgado, D., Goncalves, V., Nunes, U.: Trainable classifier-fusion schemes: An application to pedestrian detection. In: 12th International IEEE Conference on Intelligent Transportation Systems (ITSC), pp. 1–6 (2009)
36. Mansanet, J., Albiol, A., Paredes, R.: Local deep neural networks for gender recognition. Pattern Recognition Letters **70**, 80–86 (2016)
37. Ngan, M., Grother, P.: Face recognition vendor test (FRVT) performance of automated gender classification algorithms. Tech. Rep. NIST IR 8052, National Institute of Standars and Technology (2015)
38. Nixon, M., Correia, P., Nasrollahi, K., Moeslund, T., Hadid, A., Tistarelli, M.: On soft biometrics. Pattern Recognition Letters **68, Part 2**, 218–230 (2015)
39. Ojansivu, V., Heikkil, J.: Blur insensitive texture classification using local phase quantization. In: A. Elmoataz, O. Lezoray, F. Nouboud, D. Mammass (eds.) Image and Signal Processing, LNCS 5099, pp. 236–243. Springer (2008)
40. Ren, H., Li, Z.N.: Gender recognition using complexity-aware local features. In: International Conference on Pattern Recognition, pp. 2389–2394 (2014)
41. Shafey, L.E., Khoury, E., Marcel, S.: Audio-visual gender recognition in uncontrolled environment using variability modeling techniques. In: International Joint Conference on Biometrics, pp. 1 – 8 (2014)
42. Shan, C.: Learning local binary patterns for gender classification on realworld face images. Pattern Recognition Letters **33**, 431–437 (2012)
43. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: Deepface: Closing the gap to human-level performance in face verification. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1701–1708 (2014)
44. Tan, X., Triggs, B.: Enhanced local texture feature sets for face recognition under difficult lighting conditions. Image Processing, IEEE Transactions on **19**(6), 1635 – 1650 (2010)
45. Tapia, J.E., Pérez, C.A.: 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 **8**(3), 488–499 (2013)
46. Vapnik, V.: The nature of statistical learning theory. Springer, New York (1995)
47. van de Wolfshaar, J., Karaaba, M.F., Wiering, M.A.: 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)
48. Zhang, H., Beveridge, J.R., Draper, B.A., Phillips, P.J.: On the effectiveness of soft biometrics for increasing face verification rates. Computer Vision and Image Understanding **137**, 50 – 62 (2015)