Fusion of Holistic and Part Based Features for Gender Classification in the Wild

Modesto Castrillón-Santana(✉), Javier Lorenzo-Navarro, and Enrique Ramón-Balmaseda

Universidad de Las Palmas de Gran Canaria, Las Palmas de Gran Canaria, Spain
{modesto.castrillon,javier.lorenzo}@ulpgc.es
http://berlioz.dis.ulpgc.es/roc-siani

Abstract. Gender classification (GC) in the wild is an active area of current research. In this paper, we focus on the combination of a holistic state of the art approach based on features extracted from the facial pattern, with patch based approaches that focus on inner facial areas. Those regions are selected for being relevant to the human system according to the psychophysics literature: the ocular and the mouth areas. The resulting proposed GC system outperforms previous approaches, reducing the classification error of the holistic approach roughly a 30%.

Keywords: Gender classification · Local descriptors · Score level fusion

1 Introduction

Gender classification (GC) is a growing area of research with different potential applications. This fact has recently been stated by NIST in their 2015 evaluation [20]. That review highlights the difference between GC with constrained or controlled datasets, and unconstrained or *in the wild* datasets. In the first scenario, the most accurate system reached an accuracy up to 96.5% with a dataset containing almost one million samples.

However, the reported results in unconstrained imagery datasets did not present always a similar behavior. Two datasets were selected for that experiment: 1) *The Labeled Faces in the Wild* (LFW) [16], and 2) *The images of Groups* (GROUPS) [11].

Even when both datasets contain variations in terms of pose, illumination, etc., the best Face Recognition Vendor Test (FRVT) participants reported a remarkable difference in accuracy for each. For LFW, the best accuracy reached 95.2%, quite close to the numbers reported for constrained datasets. However, for GROUPS it just reached 90.4%. We can argue that this effect is due to the larger variations in terms of pose exhibited by GROUPS, and the multiple

M. Castrillón-Santana—Work partially funded by the Institute of Intelligent Systems and Numerical Applications in Engineering and the Computer Science Department at ULPGC.

V. Murino et al. (Eds.): ICIAP 2015 Workshops, LNCS 9281, pp. 43–50, 2015.
DOI: 10.1007/978-3-319-23222-5_6

samples per identity included in LFW. These results were obtained with a lights-out, black-box testing methodology.

Extending the NIST review, we summarize in Table 1 the most recent results reported in the research literature for both datasets. A fast analysis suggests that GROUPS is the most challenging one. The achieved accuracies are however not comparable to those obtained by commercial systems. The reader must observe that these results were achieved not following a lights-out, black-box testing methodology. Focusing on GROUPS, with the exception of the protocol described by Dago et al. [9], used in [4] too, the adopted protocols are not easily reproducible. The fact that GROUPS is currently the most challenging in the wild dataset, has convinced us to focus on this dataset.

Table 1. GC accuracies in recent literature. The whole dataset is used, i.e. 28000 samples, with the exception of ^{1} aropund 14000 samples with inter ocular distance > 20, ^{2} 22778 aut. detected faces, ^{3} > 12 years old, ^{4} 7443 of the total 13233 images, ^{5} BEFIT protocol, and ^{6} half dataset.

Reference	Dataset	Accuracy (%)
[9]	GROUPS^1	86.6%
[4]	GROUPS^1	89.8%
[19]	GROUPS^2	86.4%
[7]	GROUPS^2	90.4%
[3]	GROUPS^3	80.5%
[14]	GROUPS	87.1%
[23]	LFW^4	94.8%
[25]	LFW^4	98.0%
[9]	LFW^5	97.2%
[21]	LFW^6	98.0%
[3]	LFW	79.5%
[17]	LFW	96.9%
[22]	LFW	94.6%

Two recent results support the approach described in this paper. On the one hand, the extraction of features at different scales may benefit the GC performance [2,4]. In [4] the features are extracted from the face and its local context, thus, the face is analyzed at different resolutions. This fact might introduce redundancy, but the resulting improved performance suggests that an adequate design reports indeed an accuracy improvement.

On the other hand, the fusion of multiple descriptors do not just reports a benefit in GC accuracy, but also diminishes the occurrences of ambiguous cases as demonstrated for a demographics balanced dataset [6].

Therefore, the aim in this paper is to explore whether the additional integration of features extracted from specific areas of the inner face, improves the overall GC accuracy. The contributions of this work are: 1) separately the periocular and the mouth area provide an accuracy greater than 80% for GROUPS, 2) the adequate selection of periocular and mouth features, that are later fused with standard state-of-the-art facial based GC systems, provides a significant augment in terms of GC accuracy.

2 Approach

We therefore assume the ideas described above, i.e. the interest for the GC problem of a proper combination of features and regions of interest. We start from a baseline, given by a state of the art facial based GC system [4], to later explore the fusion with features densely extracted from some specific areas of the inner face [5]. With this concept in mind, we have revisited the analysis of the human visual system for the GC problem using *bubbles* [13], where the authors argue that both the ocular and the mouth areas are discriminant for this task to the human system.

An initial study of the integration of the periocular area [5] has already suggested that this approach may improve the GC performance up to 2 percentage points. Indeed, the use of components for facial analysis is a known idea. The work by Heisele et al. [15] made use of two layers of classifiers, being the second the combination of the first layer scores. The approach obtained better results than just using global features. In this paper, we indeed do not restrict to inner facial patches but also integrate features extracted from the whole facial pattern. To avoid redundancy, we select the best configuration of features, areas and grid configurations.

Summarizing, the considered patterns are presented in Figure 1: the head and shoulders (HS), the face (F), the periocular (P), and the mouth (M) areas. They all are automatically cropped from the original head and shoulders pattern (with a dimension of 155 × 159 pixels with 26 pixels of inter-eye distance), with the exception of the HS pattern that is down-sampled to 64 × 64 pixels. The original pattern is obtained after a normalization process guided by the eye locations, that encloses rotation, scaling and translation to fix the normalized eye locations.

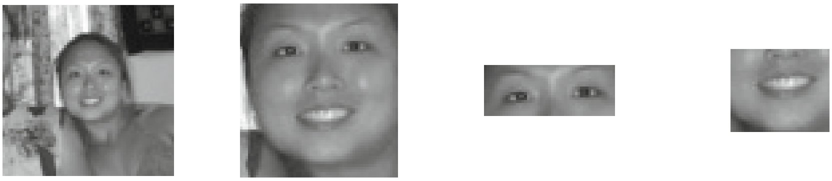

Fig. 1. From left to right, head and shoulders (HS) (64 × 64 pixels), face (F) (59 × 65 pixels), periocular (P) (37 × 31 pixels), and mouth (M) (19 × 49 pixels) regions. Sample taken from GROUPS.

After selecting the patterns to be used, we proceed with a number of steps with the final goal of evaluating the fusion or combination of multiple experts. We analyze the periocular (P) and mouth (M) areas as follows:

1. Explore the features and grid resolutions for both P and M.
2. Select the most discriminant features and grids using P and M.
3. Evaluate the combination of the state of the art GC system with the best P and M descriptors separately.
4. Evaluate the combination of the state of the art GC system with the best P and M descriptors jointly.

Based on current literature and our background related to GC, we use as features different local descriptors. Local descriptors are currently being applied for facial analysis, based on a grid configuration to avoid the loss of spatial information produced by a single based histogram representation [1].

A grid configuration is defined by its number of horizontal and vertical cells, respectively cx and cy, making a total of $cx \times cy$ cells. For a given feature, 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 by the concatenation of $cx \times cy$ histograms, i.e. $\mathbf{x} = \{h_1, h_2, ..., h_{cx \times cy}\}$.

In few words, each expert is designed with a particular feature and grid configuration. For P we have analyzed grid configurations in the range $cx \in [1, 8]$ and $cy \in [1, 6]$, while for M we have covered the range $cx \in [1, 8]$ and $cy \in [1, 8]$. That makes respectively a total of 48 and 64 variants per descriptor. As descriptors we have considered 8 different alternatives:

- Histogram of Oriented Gradients (HOG) [10].
- Local Binary Patterns (LBP) and uniform Local Binary Patterns (LBP^{u2}) [1].
- Local Gradient Patterns (LGP) [18].
- Local Ternary Patterns (LTP) [24].
- Local Phase Quantization (LPQ) [26].
- Weber Local Descriptor (WLD) [8].
- Local Oriented Statistics Information Booster (LOSIB) [12].

For the final fusion analysis, we adopt a score level fusion approach based on SVM classifiers similarly to [4,15]. The first layer is formed by the classifiers after selecting the best descriptors and grid configurations of each pattern for the problem, while the second layer classifier takes as input the first layer scores.

3 Results

As mentioned above, we adopt the Dago's protocol as experimental setup. This protocol defines a 5-fold cross validation for the GROUPS dataset. The dataset is reduced to around 14000 samples as the protocol includes only those faces that present an inter-eye distance larger than 20 pixels in the original source image.

We present in first term the results achieved making use of features extracted only from P and M. Tables 2 and 3 summarize the best results achieved for the first fold of the Dago's protocol. Due to the lack of space, only their best accuracy obtained for each descriptor configuration and pattern is included in both tables.

Table 2. Periocular based best single descriptor results in terms of accuracy (%) obtained for the first fold of the Dago's protocol. For each descriptor the best grid setup is indicated. Those descriptors providing an accuracy larger than 77% are highlighted.

	$\mathrm{HOG}_{7\times6}$	$\mathrm{LBP}^{u2}_{8\times3}$	$\mathrm{LBP}_{6\times3}$	$\mathrm{LGP}_{6\times6}$	$\mathrm{LTP}_{3\times2}$	$\mathrm{LPQ}_{2\times2}$	$\mathrm{WLD}_{6\times3}$	$\mathrm{LOSIB}_{7\times6}$
Periocular	**83.02**	**80.31**	76.24	**77.88**	**80.08**	76.00	**82.20**	76.45

Table 3. Mouth based best single descriptor results in terms of accuracy (%) obtained for the first fold of the Dago's protocol. For each descriptor the best grid setup is indicated. Those descriptors providing an accuracy larger than 75% are highlighted.

	$\mathrm{HOG}_{8\times8}$	$\mathrm{LBP}^{u2}_{5\times5}$	$\mathrm{LBP}_{4\times5}$	$\mathrm{LGP}_{7\times6}$	$\mathrm{LTP}_{3\times2}$	$\mathrm{LPQ}_{2\times2}$	$\mathrm{WLD}_{4\times5}$	$\mathrm{LOSIB}_{7\times6}$
Mouth	**80.90**	**77.71**	**74.87**	**76.62**	**77.68**	74.43	**78.23**	73.37

Table 4. Mean accuracies for the Dago's protocol with score level fusion based on the face (F), head and shoulders (HS), periocular (P) and mouth (M) areas. Each result is associated with the pattern and features fused.

Pattern(s)	Approach	Features	Acc.
P	Single	P-HOG	81.61
	Fusion	P-HOG + P-LBP^{u2} + P-LBP + P-WLD	82.79
M	Single	M-HOG	80.55
	Fusion	M-HOG + M-LBP + M-WLD + M-LGP	81.43
F+HS	Fusion	F-HOG + F-LBP^{u2} + HS-HOG	**90.49**
F+HS+P	Fusion	F-HOG + F-LBP^{u2} + HS-HOG P-HOG + P-LBP^{u2} + P-LBP + P-WLD + P-LOSIB	92.42
F+HS+M	Fusion	F-HOG + F-LBP^{u2} + HS-HOG M-HOG + M-LBP + M-WLD + M-LGP	91.60
F+HS+P+M	Fusion 1	F-HOG + F-LBP^{u2} + HS-HOG P-HOG + P-LBP^{u2} + M-HOG + M-WLD	**93.22**
	Fusion 2	F-HOG + F-LBP^{u2} + HS-HOG P-HOG + P-LBP^{u2} + P-LGP + M-HOG	93.22
	Fusion 3	F-HOG + F-LBP^{u2} + HS-HOG P-HOG + P-LBP^{u2} + M-LGP + M-HOG	93.15

The accuracies are slightly worse for M compared to P. Being in both cases significantly worse than those reported by recent face based GC systems. For the later fusion analysis, we have selected those descriptors providing an accuracy larger than 77% for P, and larger than 75% for M. Making a total of 11 descriptors.

The next step considers the fusion of the most discriminant descriptor setups with the state of the art approach described in [4]. This approach extracts HOG and LBP features from F (F-HOG and F-LBP), and HOG from HS (HS-HOG). The fusion is evaluated first separately with the best descriptors for P and M, including exhaustive search among all possible combinations. This means that we evaluated all possible combinations with P, 2^5, and M, 2^6.

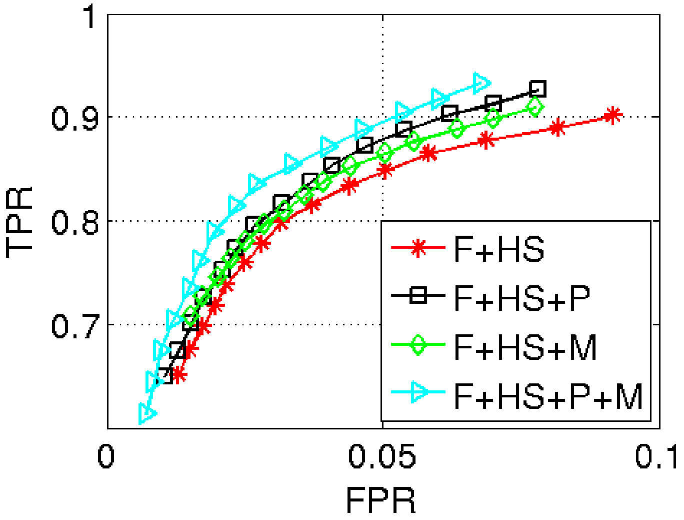

Fig. 2. ROC curves using the Dago's protocol. Comparison of state-of-the-art classification based on F and HS, with the proposed fusion alternatives considering HS and F features respectively with P, M and both.

The final experiment evaluates the fusion with both sets of descriptors, i.e. covering the whole range of possible combinations, i.e. 2^{11} possibilities. The results reported for each approach are summarized in Table 4 indicating the best descriptors combination. 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 $gamma = [0.04, 0.15]$.

As suggested by the table, the best approach fuses descriptors extracted from all the patterns. We have included the top-3 approaches, they report quite similar accuracies, but certainly those using a lower number of features will reduce the processing cost.

A detailed observation indicates that for the Dago's protocol, the improvement in accuracy is close to 3 percentage points. Observing the resulting ROC curves (only the best results for each fusion approach is presented in Figure 2) the fusion with P alone, is better than fusing with M. However, the combination with both P and M reports better performance, in terms of accuracy and AUC, that not using features extracted from any of those inner facial areas.

4 Conclusions

In this paper, we have explored the benefits of combining holistic features with features extracted from specific inner facial regions. In particular we have focused on the ocular and mouth areas, that have evidenced their main importance for this task in the human system.

The achieved results indicate a promising line of research. Indeed the GC performance increased up to 3 percentage points, reducing the gap present in GC accuracy with other simpler datasets. Observing the error for the facial based state of the art approach, 9.5%, the proposed systems reduces the gender classification error in more than 28%.

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), December 2006
2. Alexandre, L.A.: Gender recognition: A multiscale decision fusion approach. Pattern Recognition Letters **31**(11), 1422–1427 (2010)
3. Bekios-Calfa, J., Buenaposada, J.M., Baumela, L.: Robust gender recognition by exploiting facial attributes dependencies. Pattern Recognition Letters **36**, 228–234 (2014)
4. Castrillón-Santana, M., Lorenzo-Navarro, J., Ramón-Balmaseda, E.: Improving gender classification accuracy in the wild. In: Ruiz-Shulcloper, J., Sanniti di Baja, G. (eds.) CIARP 2013, Part II. LNCS, vol. 8259, pp. 270–277. Springer, Heidelberg (2013)
5. Castrillón-Santana, M., Lorenzo-Navarro, J., Ramn-Balmaseda, E.: Evaluation of periocular over face gender classification in the wild (under review)
6. Castrillón-Santana, M., Marsico, M.D., Nappi, M., Riccio, D.: MEG: Multi-Expert Gender classification in a demographics-balanced dataset. In: 18th International Conference on Image Analysis and Processing (2015)
7. 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)
8. Chen, J., Shan, S., He, C., Zhao, G., Pietikainen, M., Chen, X., Gao, W.: WLD: A robust local image descriptor. IEEE Transactions on Pattern Analysis and Machine Intelligence **32**(9), 1705–1720 (2010)
9. 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 (2011)
10. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Schmid, C., Soatto, S., Tomasi, C. (eds.) International Conference on Computer Vision & Pattern Recognition, vol. 2, pp. 886–893, June 2005
11. Gallagher, A., Chen, T.: Understanding images of groups of people. In: Proc. CVPR (2009)
12. García-Olalla, O., Alegre, E., Fernández-Robles, L., González-Castro, V.: Local oriented statistics information booster (LOSIB) for texture classification. In: International Conference in Pattern Recognition (ICPR) (2014)
13. Gosselin, F., Schyns, P.G.: Bubbles: a technique to reveal the use of information in recognition tasks. Vision Research, 2261–2271 (2001)
14. Han, H., Jain, A.K.: Age, gender and race estimation from unconstrained face images. Tech. Rep. MSU-CSE-14-5. Michigan State University (2014)
15. Heisele, B., Serre, T., Poggio, T.: A component-based framework for face detection and identification. International Journal of Computer Vision Research **74**(2), August 2007

16. 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, October 2007
17. Jia, S., Cristianini, N.: Learning to classify gender from four million images. Pattern Recognition Letters (2015)
18. Jun, B., Kim, D.: Robust face detection using local gradient patterns and evidence accumulation. Pattern Recognition **45**(9), 3304–3316 (2012)
19. 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 (PAMI), October 2011
20. Ngan, M., Grother, P.: Face recognition vendor test (frvt) performance of automated gender classification algorithms. Tech. Rep. NIST IR 8052. Narional Institute of Standars and Technology, April 2015
21. Ren, H., Li, Z.N.: Gender recognition using complexity-aware local features. In: International Conference on Pattern Recognition (2014)
22. Shafey, L.E., Khoury, E., Marcel, S.: Audio-visual gender recognition in uncontrolled environment using variability modeling techniques. In: International Joint Conference on Biometrics (2014)
23. Shan, C.: Learning local binary patterns for gender classification on realworld face images. Pattern Recognition Letters **33**, 431–437 (2012)
24. Tan, X., Triggs, B.: Enhanced local texture feature sets for face recognition under difficult lighting conditions. IEEE Transactions on Image Processing **19**(6), 1635–1650 (2010)
25. 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)
26. Ojansivu, V., Heikkilä, J.: Blur insensitive texture classification using local phase quantization. In: Elmoataz, A., Lezoray, O., Nouboud, F., Mammass, D. (eds.) ICISP 2008 2008. LNCS, vol. 5099, pp. 236–243. Springer, Heidelberg (2008)