Improving Gender Classification Accuracy in the Wild

Modesto Castrillón-Santana*, Javier Lorenzo-Navarro, and Enrique Ramón-Balmaseda

SIANI
Universidad de Las Palmas de Gran Canaria, Spain
mcastrillon@siani.es

Abstract. In this paper, we focus on gender recognition in challenging large scale scenarios. Firstly, we review the literature results achieved for the problem in large datasets, and select the currently hardest dataset: The Images of Groups. Secondly, we study the extraction of features from the face and its local context to improve the recognition accuracy. Different descriptors, resolutions and classifiers are studied, overcoming previous literature results, reaching an accuracy of 89.8%.

Keywords: gender recognition, local context, head and shoulders, LBP, HOG, in the wild.

1 Introduction

Gender is a valid demographic characteristic for different applications that has recently attracted commercial attention in the context of audience analysis and advertisement.

Different approaches have tackled the problem of automatic gender recognition. Most recent works have basically considered the face pattern to solve the problem [2, 3, 14]. Other approaches have made use of non facial features such as the whole body, the hair or clothing [4, 13]. However, those approaches including non facial features, have rarely considered uncontrolled large datasets, i.e. the gender recognition in the wild. In this context, the evaluation must tackle more variability in terms of 1) identities, aging and ethnicity, 2) pose and illumination control, and 3) low resolution images.

The contributions of this paper rely firstly on the addition to the information provided by the face, of features extracted from the head local context. Those features are studied at different resolutions, and their possibilities analyzed as additional features for the problem. Another main element of this paper is the use of large databases that are closer to real gender classification scenarios than those small databases obtained in controlled environments.

* This work was partially funded by the Institute of Intelligent Systems and Numerical Applications in Engineering (SIANI) and the Computer Science Department at ULPGC.

1.1 Previous Work in Large Datasets

Table 1. Gender recognition accuracy in the previous literature. Refer to each reference for experimental setup details.

Reference	Dataset	Protocol	Accuracy
[19]	LFW	Subset 7443/13233	94.81%
[20]	LFW	Subset 7443/13233	98.01%
[7]	LFW	BEFIT protocol	97.23%
[7]	GROUPS	Subset 15579/28231	$84.55 - 86.61\%$
[12]	GROUPS	Subset 22778/28231	86.4%
[5]	MORPH	Subset	88%
[17]	MORPH	Subset	97.1%

We argue that small databases are not representative for a real world scenario where a gender recognizer must cope with thousands of people, like for example a mall scenario. For that reason, we have reviewed the literature to detect state-of-the-art accuracies obtained for large public databases that contain many different identities acquired without controlled conditions. As far as we know, Table 1 presents the best accuracies reported on large datasets in the recent literature. The datasets studied are The Image of Groups (GROUPS) [10], Labeled Faces in the Wild (LFW) [11], and MORPH [18].

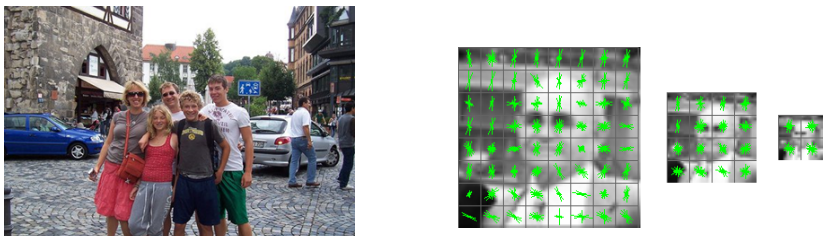
Observing in detail Table 1, there is not much space for improvement in datasets such as LFW and MORPH. Certainly, both datasets present a set of characteristics that might affect the impressive resulting performance. Indeed, in both datasets the same identity includes multiple samples. Additionally, sample images of both genders are not equally represented in the set, i.e. the number of samples corresponding to the male class is significantly larger. On the other side, the GROUPS dataset presents unrestricted imagery with balanced presence of both classes, reporting the lowest accuracy in the recent works. For all those reasons, we have selected to focus on the GROUPS dataset, that represents, in our opinion, the wildest available dataset for the problem, see Figure 1a.

2 Representation and Classification

Local descriptors have recently attracted the attention of researchers involved in the facial analysis community [21]. We will focus particularly on Local Binary Patterns [16] (LBPs) and Histograms of Oriented Gradients [8]. Both descriptors have already been used successfully for facial analysis [9, 15].

Facial analysis with LBP is currently adopted considering a concatenation of histograms of a predefined grid. This approach was adopted for LBP by Ahonen et al. [1]. According to that work the face is divided into regions where the LBP operator is computed and later their corresponding histograms concatenated, following a Bag of Words scheme [6], into a single histogram. On the other side,

HOG encloses a histogram in its definition. The pattern is scaled to a normalized resolution, and later a grid is defined.



(a) (b)

Fig. 1. a) A GROUPS sample. b) Relative size of the different patterns used including the local context: respectively 64×64, 32×32 and 16×16 faces with head and shoulders context. Their respective HOG grid computed is also depicted.

For classification purposes, we will compare two different approaches. The first one will study the addition of features to an initial feature vector filled exclusively with features extracted from the face. In this scenario, two well known classifiers are compared: SVM with linear kernel, and bagging making use of SVM classifiers based on linear kernels.

In the second scenario, instead of combining features of different nature in the feature vector, we focus on the combination of the outputs provided by the different classifiers in a first stage. Their respective scores are combined in a second classification stage. This combination is compared based on different known classification techniques such as: SVM (linear kernel), bagging, naive Bayes, Nearest Neighbor (NN) and C4.5.

3 Results

In the experimental setup, we have adopted a k-fold cross-validation, partitioning the dataset into k subsets, repeating k times the experiment using a subset to test the model with the other $k-1$ subsets. In order to be comparable to previous works, we made use of the 5-folds defined in the work by Dago et al. [7].

The Uniform LBP descriptor is used only for the face area, at two different resolutions: 59×65 and 100×110, defining a 5×5 grid. When using HOG as descriptor, the face area is used just with the 59×65 resolution, but the head and shoulders pattern was tested at different resolutions: 16×16, 32×32 and 64×64, see Figure 1b. On each resolution the cell contains 8×8 pixels, each block 2×2 cells, the histogram contains 9 bins, and L2-hys as norm for the normalization stage [8].

3.1 Extending the feature vector

Firstly, we performed a comparison using just the facial information, i.e. the inner face details (Face), and its local context defined by the head and shoulders area (HS). Tables 2 and 3 present respectively those results. The face pattern resolution used in Table 2 was 59×65 pixels, with an inter-eye distance of 26 pixels. For comparison with a baseline, we have also included the results achieved with a classifier trained with the first 100 PCA components.

Table 2. Gender recognition accuracy (in brackets results per class: female/male) achieved using PCA, HOG or Uniform LBP features extracted from the face pattern. The table reports the results achieved using the 5-fold cross correlation experiment defined by Dago et al. SVM linear and Bagging are used for classification.

Pattern and features	Test set GROUPS-Dago 5-folds subset SVM linear Acc.	AUC	Bagging Acc.	AUC
Face PCA	0.773 (0.773/0.774)	0.773	0.7749 (0.779/0.770)	0.801
Face HOG	0.801 (0.797/0.805)	0.801	0.822 (0.84/0.800)	0.898
Face LBP	**0.838** (0.842/0.834)	0.838	**0.838** (0.863/0.814)	0.910

Table 3 presents the results using information extracted from the face and its local context. The head and shoulders were analyzed at different resolutions: 16×16, 32×32 and 64×64, with their respective inter-eye distances of 2.5, 5 and 10 pixels. Observe, that the facial resolution contained in the head and shoulders pattern is lower up to ten times compared to the results reported in Table 2. Even though, the best accuracy is rather similar, almost 84% using the 64×64 head and shoulders pattern, than exclusively the facial pattern at larger resolution. Even considering the smallest pattern, with an inter-eye distance under 3 pixels, the accuracy reaches 66%. That is not a bad result for low resolution images.

On a second step, we have considered to fuse in a single feature vector, features extracted from different cues. Table 4 presents results combining Uniform LBP or HOG features extracted from the face pattern, with HOG features extracted from the head and shoulders pattern at different resolutions. Bagging reports better accuracy for the experimental setup, while the use of Uniform LBP features seems to work slightly better than HOG. The notorious increase in the face pattern resolution, does not suggest a large improvement in accuracy. The best reported accuracy reaches 88.1%, four points better than our previous results, and 2% better than the literature for the same dataset, see Table 1. These results suggest the importance of the information contained in the facial local context.

Table 3. Gender recognition accuracy (in brackets results per class: female/male) achieved using HOG features extracted from the head and shoulders (HS) pattern using different image dimensions. The table reports the results achieved using the 5-fold cross correlation experiment defined by Dago et al. SVM linear and Bagging are used for classification.

Pattern and features	Test set GROUPS-Dago 5-folds subset			
	SVM linear		Bagging	
	Acc.	AUC	Acc.	AUC
$HS_{16\times16}$ HOG	0.6608 (0.6538/0.6684)	0.661	0.659 (0.6616/0.6564)	0.687
$HS_{32\times32}$ HOG	0.812 (0.8024/0.8216)	0.812	0.8099 (0.8122/0.8076)	0.865
$HS_{64\times64}$ HOG	0.8298 (0.829/0.83)	0.829	**0.8397** (0.8562/0.8232)	0.909

3.2 Stacking Classifiers

We went further, and considered an alternative to the inclusion of more features in the feature vector. Instead, we considered a stacking of classifiers in two stages. The first stage is composed by the individual 11 feature vectors described in Tables 3 and 4, and summarized in the following list:

- HOG of the 64×64 head and shoulders pattern (HSHOG64).
- HOG of the 32×32 head and shoulders pattern (HSHOG32).
- HOG of the 16×16 head and shoulders pattern (HSHOG16).
- HOG of the 59×65 facial pattern (FHOG).
- Concatenated LBP histogram extracted from the 59×65 facial pattern (FLBP).
- HOG of the 64×64 head and shoulders pattern, and HOG of the 59×65 face pattern (HSHOG64-FHOG).
- HOG of the 32×32 head and shoulders pattern, and HOG of the 59×65 face pattern (HSHOG32-FHOG).
- HOG of the 16×16 head and shoulders pattern, and HOG of the 59×65 face pattern (HSHOG16-FHOG).
- HOG of the 64×64 head and shoulders pattern, and concatenated LBP histogram of the 59×65 face pattern (HSHOG64-FLBP).
- HOG of the 32×32 head and shoulders pattern, and concatenated LBP histogram of the 59×65 face pattern (HSHOG32-FLBP).
- HOG of the 16×16 head and shoulders pattern, and concatenated LBP histogram of the 59×65 face pattern (HSHOG16-FLBP).

Each of the first stage classifier is trained using a SVM with a liner kernel. In the second stage of the stacking classifier, their respective scores are feed into a classifier that is in charge of taking the final decision. For this second stage, we have analyzed the accuracy reported for SVM (linear kernel), Bagging, Naive Bayes, Nearest Neighbor (NN) and C4.5. The results achieved are reported in

Table 4. Gender recognition accuracy (in brackets results per class: female/male) achieved using different representation alternatives. The table reports the results achieved using the 5-fold cross correlation experiment defined by Dago et al. SVM linear and Bagging are used for classification.

	Test set GROUPS-Dago 5-folds subset							
	SVM linear				Bagging			
Pattern and	face 59 × 65		face 100 × 110		face 59 × 65		face 100 × 110	
features	Acc.	AUC	Acc.	AUC	Acc.	AUC	Acc.	AUC
Face HOG	0.801 (0.797/0.805)	0.801	-	-	0.822 (0.84/0.800)	0.898	-	-
Face LBP	0.827 (0.835/0.814)	0.827	0.836 (0.836/0.836)	0.836	0.823 (0.856/0.804)	0.905	0.84 (0.862/0.817)	0.909
Face HOG and $HS_{16\times16}$ HOG	0.827 (0.828/0.827)	0.827	-	-	0.829 (0.848/0.809)	0.904	-	-
Face HOG and $HS_{32\times32}$ HOG	0.852 (0.855/0.85)	0.852	-	-	0.862 (0.881/0.843)	0.93	-	-
Face HOG and $HS_{64\times64}$ HOG	0.845 (0.851/0.84)	0.845	-	-	0.875 (0.893/0.858)	0.941	-	-
Face LBP and $HS_{16\times16}$ HOG	0.838 (0.842/0.834)	0.838	0.844 (0.843/0.846)	0.844	0.838 (0.863/0.814)	0.910	0.845 (0.864/0.826)	0.915
Face LBP and $HS_{32\times32}$ HOG	0.859 (0.86/0.857)	0.859	0.862 (0.861/0.863)	0862	0.867 (0.889/0.845)	0.933	0.869 (0.864/0.826)	0.937
Face LBP and $HS_{64\times64}$ HOG	0.851 (0.851/0.85)	0.851	0.861 (0.859/0.864)	0.861	0.879 (0.897/0.862)	0.944	**0.881** (0.897/0.866)	**0.946**

Table 5. They suggest an improvement, reaching with Naive Bayes almost 90%. The reader may observe, that this accuracy was achieved without using the classifiers based on the largest face pattern, i.e. an inter-eye distance of 26 pixels. Compared to Table 4 for similar facial resolution the improvement is almost 2%. Compared to the literature, see Table 1, the improvement is close to 4% even using a facial pattern that is twice smaller. The benefits introduced by the descriptors and the face local context are evident.

We have additionally performed a feature selection to reduce the system complexity avoiding the computation of all the classifiers present in the stacking first stage. After sorting attending to the information gain, the resulting accuracy considering as variable the number of classifiers included in the stacking is presented in Figure 2. With just 4 classifiers in the first stage, the system performance achieves an accuracy of 89% beating those results reported in the previous section and the literature. Those classifiers are: HSHOG64-FLBP, HSHOG32-FLBP, HSHOG16-FLBP and HSHOG32-FHOG.

Table 5. Gender recognition accuracy achieved using classifiers stacking. The table reports the results achieved using the 5-fold cross correlation experiment defined by Dago et al.

Classifier	Accuracy
Naive Bayes	**0.8978**
SVM	0.8736
C4.5	0.8336
NN	0.8662

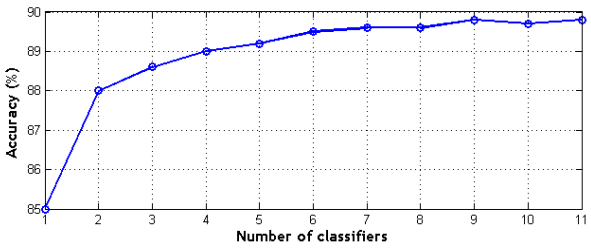


Fig. 2. Accuracy achieved adding more classifiers to the stacking.

4 Conclusions

In this paper, we have studied gender recognition in large uncontrolled datasets. For that purpose, we have made use of facial and non facial features, in the large database that is currently reporting the lowest accuracy in the literature: The Images of Groups.

The addition of external facial features seem to bring benefits at lower resolution, and the combination with facial features reported better accuracies that the previous literature.

We have used features based on the Uniform LBP and HOG operators, both used widely in similar problems. For classification we have considered the used combination of features in a large dataset, and the stacking of classifiers, each one focused in a particular family of features. The stacking results are particularly better than those obtained when the feature vector is increased, reaching almost 90%. This accuracy is notoriously better than those previously reported in the literature, even if the face pattern considered makes use of a facial resolution at least twice smaller.

In summary, the performance exhibited at lower resolution, is best suited for real scenarios. However, the achieved at high resolution beats state of the art results.

References

1. 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*, 28(12), December 2006.
2. Luis A. Alexandre. Gender recognition: A multiscale decision fusion approach. *Pattern Recognition Letters*, 31(11):1422–1427, 2010.
3. Juan Bekios-Calfa, José M. Buenaposada, and Luis Baumela. Revisiting linear discriminant techniques in gender recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(4):858–864, April 2011.
4. Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Describing people: A poselet-based approach to attribute classification. In *International Conference on Computer Vision*, 2011.

5. 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.
6. Gabriella Csurka, Christopher R. Dance, Lixin Fan, Jutta Willamowski, and Cdric Bray. Visual categorization with bags of keypoints. In *Workshop on Statistical Learning in Computer Vision, ECCV*, pages 1–22, 2004.
7. P. Dago-Casas, D. González-Jiménez, L. Long-Yu, and J. L. 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.
8. Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In Cordelia Schmid, Stefano Soatto, and Carlo Tomasi, editors, *International Conference on Computer Vision & Pattern Recognition*, volume 2, pages 886–893, INRIA Rhône-Alpes, ZIRST-655, av. de l'Europe, Montbonnot-38334, June 2005.
9. O Déniz, G Bueno, J Salido, and F De La Torre. Face recognition using histograms of oriented gradients. *Pattern Recognition Letters*, 32(12):1598–1603, 2011.
10. A. Gallagher and T. Chen. Understanding images of groups of people. In *Proc. CVPR*, 2009.
11. 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. Technical Report 07-49, Univ, of Massachusetts, Amherst, October 2007.
12. 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 (PAMI)*, October 2011.
13. Bing Li, Xiao-Chen Lian, and Bao-Liang Lu. Gender classification by combining clothing, hair and facial component classifiers. *Neurocomputing*, 76(1):18–27, January 2012.
14. Erno Mäkinen and Roope Raisamo. Evaluation of gender classification methods with automatically detected and aligned faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(3):541–547, March 2008.
15. Sébastien Marcel, Yann Rodríguez, and Guillaume Heusch. On the recent use of local binary patterns for face authentication. *International Journal of Image and Video Preprocessing, Special Issue on Facial Image Processing*, 2007.
16. 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*, 24(7):971–987, 2002.
17. E. Ramón-Balmaseda, J. Lorenzo-Navarro, and M. Castrillón-Santana. Gender classification in large databases. In *17th Iberoamerican Congress on Pattern Recognition (CIARP)*, 2012.
18. 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*, pages 341–345., Southampton, UK, April 2006.
19. Caifeng Shan. Learning local binary patterns for gender classification on realworld face images. *Pattern Recognition Letters*, 33:431437, 2012.
20. Juan E. Tapia and Claudio A. Perez. 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.
21. Lior Wolf, Tal Hassner, and Yaniv Taigman. Descriptor based methods on the wild. In *In: Faces in Real-Life Images Workshop in ECCV*, 2008.