A performance based study on gender recognition in large datasets

M. Díaz-Cabrera, J. Lorenzo-Navarro, and M. Castrillón-Santana*

SIANI
Universidad de Las Palmas de Gran Canaria Spain
moisdc@gmail.com
{jlorenzo,mcastrillon}@siani.es
http://www.siani.es

Abstract. Gender recognition has achieved impressive results based on the face appearance in controlled datasets. Its application in the wild and large datasets is still a challenging task for researchers. In this paper, we make use of classical techniques to analyze their performance in controlled and uncontrolled condition respectively with the LFW and MORPH datasets. For both sets the benchmarking protocol follows the 5-fold cross-validation proposed by the BEFIT challenge.

Keywords: Gender recognition, BEFIT, classifier fusion, LFW, MORPH

1 Introduction

Gender is, among other static and dynamic features, easily extracted and integrated during interaction by humans. Gender classification is indeed a research field in computer vision, with different application scenarios covering demographics or direct marketing among others.

Automatic gender classification is therefore an active topic as evidenced by recent publications in major journals [2, 15, 17]. Nowadays state-of-the-art approaches are based on the facial appearance. The interested reader should consider the work by Mäkinen et al. [15] a valuable source presenting a comparison of classification results for this problem with automatically detected faces.

Until very recent papers, comparison results have been mostly presented for the FERET [19] database. This database provides images of good quality of a relatively reduced number of individuals. Nowadays, there are other databases exhibiting much larger variations in terms of 1) identities, aging and ethnicity, and 2) pose and illumination control. Among the first group we can mention the results reported in [2, 5, 6] respectively for the non public UCN, the MORPH [12], and the Gallager's [8] databases (all of them containing more than 10000 different identities). Those results achieved for large databases do not agree completely with previous studies [2]. In the second group, we should mention the LFW [11]

* Work partially funded by the Spanish Ministry of Science and Innovation funds (TIN 2008-06068), and the Computer Science Department at ULPGC.

database containing multiple images of famous people in uncontrolled imagery, i.e. faces in the wild.

In this paper we will focus on gender classification applied to the selected databases of the BEFIT benchmarking protocol [7]: the LFW and the MORPH databases revisiting classical techniques to illustrate researchers the current challenges on this topic.

2 Face space representation

Face images are highly dimensional. To reduce the problem dimension, Principal Component Analysis (PCA) is a classical method employed without a significant loss of information [13]. A normalized face image is projected in the PCA space. The face appearance is then represented in a space of lower dimensionality by means of a number of the resulting coefficients [23].

Local descriptors have received lots of attention recently in the context of facial analysis [4, 25]. Among them, Local Binary Patterns (LBPs) have been extensively used for facial recognition [1, 16], facial expression [21], demographics [26], etc.

The LBPs were originally introduced for texture classification [18], presenting invariance to monotonic changes in gray scale and low processing cost. The LBP operator analyzes the circular pixel neighborhood within a distance in the gray image, labeling the center pixel considering the result as a binary pattern. The basic version, considers the eight neighbors of the center pixel, but its definition can be extended to any radius, R, considering P neighbors [18]:

$$LBP_{P,R}(x_c, y_c) = \sum_{k=0}^{p-1} s(g_p - g_c)2^k \ , \ s(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases} \tag{1}$$

where g_c the gray level of the center pixel, and g_p the gray level of the p-neighbor.

Ojala et al. observed that only a reduced subset of local binary patterns, known as uniform. These patterns are characterized by the fact that they contain, at most, two bitwise transitions from 0 to 1 or viceversa. For example, 00000000, 00011100 and 10000011 are uniform patterns. They cover more than 90% of all patterns when using the 3x3 neighborhood. Considering the LBP pattern as circular, rotation invariance is achieved.

To represent the facial appearance, LBPs can be applied as preprocessing emphasizing edges and noise, see figure 1. But different authors have used the resulting histogram based to describe the image [16]. However, this approach can present the risk of losing relative location information [16, 22]. For that reason, when histograms are considered, a grid representation is typically used [20, 26]. Both in LBP histogram based approach and preprocessing, PCA can be applied to reduce the dimensionality of the problem.

3 Datasets

Recently, different initiatives have made public databases containing challenging situations without controlled acquisition parameters, or including large variations in terms of identities, aging and ethnicity. The BeFIT protocol [7] has defined the experimental setup for two databases: the LFW [11] and the MORPH [12] databases. The first one focuses on uncontrolled imagery, while the second on the variations in ethnicity, age, and identity.

To the best of our knowledge, only two works have attempted the problem for the LFW database and one for the MORPH. A recent attempt on the LFW database [20] reaches 94% using a 5-folds but excluding the non frontal images, i.e. reducing the uncontrolled condition of the dataset. The experiment studies the performace not in the original set of more than 13000 facial images, but in around 7000. No details are given related to the risk of using faces of the same individual in different folds. That risk is avoided in the BeFIT gender recognition challenge [7]. In [6] the normalized images used for the experiments present 45 pixels of intra-ocular distance, to evaluate classical techniques following the BEFIT protocol. The results reported achieved an accuracy of 93.5% based on a Gabor jets representation, while 88.3% for the larger Gallagher's database [8].

For the larger MORPH database, we have not found any other reference than the work described in [5]. In that paper the reported accuracy is lower than 88%, but not precise details are given about the k-fold configuration in the experimental setup.

For our experiments, face images have been annotated based on a semi-automatic approach (automatic for 85% of the images) based on the Viola-Jones based detection [24] of the face [14] and eyes [3]. The rough eye information is then used to scale and crop each image to 59 × 65 pixels, see for example figure 1b.

4 Experiments

4.1 Metrics

To evaluate the classification results, we made use of common metrics, considering in alphabetical order the female class as negative (class 0), and the male class as positive (class 1).

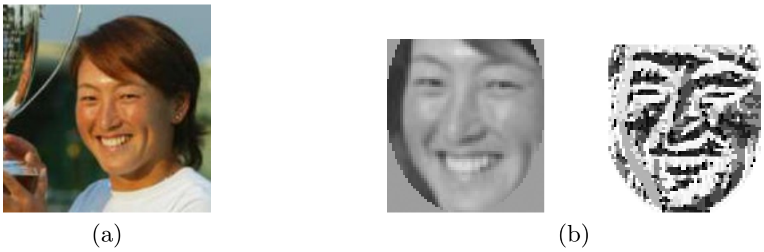
$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \tag{2}$$

$$Recall\ or\ TPR = \frac{TP}{TP+FN} \tag{3}$$

$$TNR = \frac{TN}{TN+FP} \tag{4}$$

where TP, TN, FP and FN are respectively true positives, true negatives, false positives and false negatives. We include also the Area under the Receiver

Operator Characteristic Curve (AUC), as it is commonly used to evaluate different classifiers.



(a)

(b)

Fig. 1. (a) LFW sample, (b) normalized LFW sample before and after LBP preprocessing.

4.2 Gender classification

To evaluate each classifier, we have followed the 5-fold cross-validation protocol defined for the BeFIT gender recognition challenge for both datasets. To give an impression of both databases, table 1 indicates the number of samples per class contained in each of the given 5 folds. It is evident that the number of female samples is lower in both datasets, a clear protocol drawback.

Table 1. Samples per class in each fold.

LFW				
	Train		Test	
Fold	**Male**	**Female**	**Male**	**Female**
0-1	8204	2381	2052	596
2-4	8204	2382	2051	595

MORPH				
	Train		Test	
Fold	**Male**	**Female**	**Male**	**Female**
0	37296	6788	9345	1700
1	37334	6762	9307	1726
2	37324	6794	9317	1694
3	37330	6774	9311	1714
4	37280	6834	9361	1654

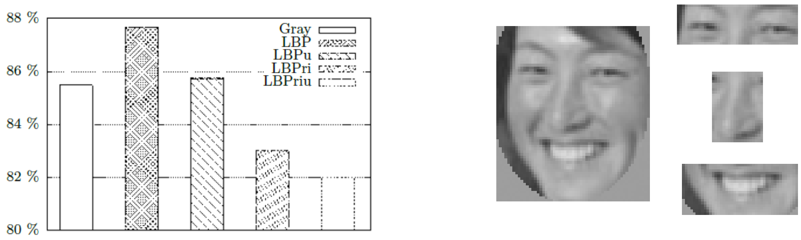
4.3 LFW

For the LFW dataset, the collection of classifiers tested cover different face representation spaces:

- the original gray values (Gray),
- preprocessed (no histogram based representation) original LBP (LBP),
- uniform LBP (LBPu),

- rotation invariant LBP (LBPri), and
- uniform rotation invariant LBP (LBPriu).

For each input, we have compared with Weka [10] the performance achieved using the pixels values (gray or preprocessed) and their PCA projection. In our experiments, the use of PCA reported always better recognition rates. Due to the space restrictions, only those results based on the PCA projection are included in this paper. Once assumed the use of PCA based representations, we focused on studying the use of the different input sources mentioned above.



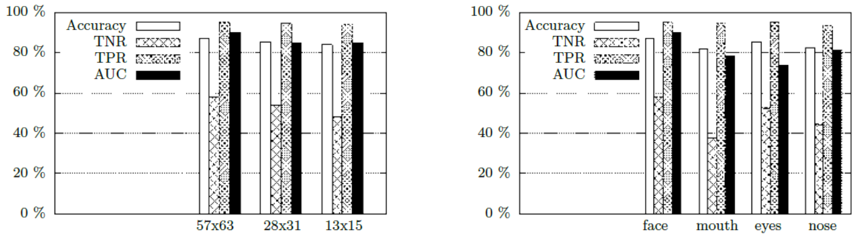
(a) (b)

Fig. 2. (a) Accuracy average for Gray, LBP, LBPu, LBPri, LBPriu using 50 PCA components. Image size 57 × 63 pixels (b) Normalized face and selected face parts: eyes, nose and mouth.

Table 2. Metrics achieved per class and experiment using linear SVM based on the normalized face with 50 components.

Fold	**TPR**	**TNR**	**Accuracy**	**AUC**
0	0.96	0.63	0.88	0.92
1	0.95	0.52	0.85	0.87
2	0.95	0.59	0.87	0.89
3	0.95	0.59	0.87	0.88
4	0.95	0.58	0.86	0.88
Mean	0.95	0.58	0.87	0.9

Figure 2a presents the mean accuracy obtained in the LFW 5-fold experiment for each representation space: Gray, LBP, LBPu, LBPri, LBPriu. The best results, employing just 50 PCA components, were obtained for the LBP input images using a linear SVM classifier. Table 2 reflects the metrics achieved for each folder using this classifier. The bias observed for the TPR and TNR is likely due to the unbalanced training set. The LBP representation based approach is used in the rest of the study for the LFW database.



(a) (b)

Fig. 3. Accuracy, sensitivity (TPR), specificity (TNR) and AUC for using 50 PCA components (a) different face sizes and (b) for the face, eyes, mouth and nose.

The influence of the pattern dimension is depicted in Figure 3a. As suggested by previous studies [15, 17], the performance decrease is not dramatic even with low resolutions images.

A psychophysical study on gender classification in humans based on the ”bubbles” technique [9], suggested that the gender can be correctly determined by humans using just the eyes and mouth areas. Thus, we have also studied inner areas of the normalized face for this problem, see the selected areas shown in figure 2b.

Figure 3b presents the mean results, making use of the face compared with three different face areas: eyes, mouth and nose. Despite the face pattern seems to include more information, the results are just slightly worse for the eyes pattern, and not extremely bad for the other two patterns. For the face and eyes, the SVM classifier made use of linear kernels, while the radial kernel was used for the nose and mouth. The reader can observe that the latter two patterns report particularly worse results for the female class (TNR).

To conclude the study, we have analyzed the influence of the number of PCA components used to represent the face space. In the comparison presented in table 3, the reader can observe that the use of 100 PCA components improves the accuracy for the face pattern, slightly for the eyes pattern, and there is no a real improvement for the nose and mouth patterns.

Table 3. Accuracy comparison for each pattern using 50 and 100 components.

	Accuracy							
	face		eyes		nose		mouth	
Fold	50	100	50	100	50	100	50	100
0	0.885	0.898	0.880	0.879	0.844	0.847	0.837	0.833
1	0.850	0.861	0.847	0.853	0.810	0.810	0.796	0.799
2	0.873	0.888	0.866	0.873	0.834	0.833	0.831	0.826
3	0.869	0.883	0.857	0.866	0.814	0.815	0.818	0.813
4	0.864	0.876	0.864	0.866	0.825	0.829	0.813	0.817
Mean	0.868	**0.881**	0.863	**0.867**	0.825	**0.826**	**0.819**	0.818

4.4 MORPH

For the MORPH database we have restricted the study to the following two input sources: Gray and LBP. Both sources are projected to their respective PCA spaces. The normalized face dimensions was respectively 59×65 and 57×63 pixels, and the number of PCA components fixed to 100.

Attending to table 4, the results achieved for the gray input are even slightly better than those obtained using LBP. This database is characterized by having almost 5 times more images than LFW, and a more reduced fraction of female images present. For this larger database the use of the gray levels as input reported better results than LBP. Table 4 presents the results for both representation spaces using the stacking approach. This time the radial SVM have been used for each SVM classifier.

Table 4. Accuracy results achieved for the MORPH database using gray or LBP as input, and projected to a PCA space.

LBP				
Fold	Face	Eyes	Nose	Mouth
0	0.91	0.88	0.88	0.89
1	0.91	0.88	0.88	0.88
2	0.91	0.88	0.88	0.89
3	0.91	0.88	0.88	0.89
4	0.91	0.88	0.88	0.89
mean	0.91	0.88	0.88	0.89

Gray				
Fold	Face	Eyes	Nose	Mouth
0	0.91	0.89	0.88	0.91
1	0.91	0.89	0.88	0.91
2	0.92	0.90	0.88	0.92
3	0.92	0.89	0.88	0.92
4	0.92	0.89	0.89	0.92
mean	0.92	0.89	0.88	0.92

4.5 Stacking based gender classification

In the previous sections we have observed that the LBP preprocessed image provides better results for the LFW database, but slightly worse than the gray image for the MORPH database. However, a closer look suggests that female samples are frequently wrongly classified. This might be justified by the unbalanced datasets. Additionally, if a specific section of the face area is analyzed, the accuracy results are also over 80% for both databases.

In this section, we will combine the resulting probabilities provided by different classifiers in a stacking fashion. The level-1 model is a linear/radial SVM classifier that is feed with the probability estimated by the base models.

Tables 5 and 6 present the results combining the outputs of the four classifiers (face, eyes, mouth and nose). The results are slightly better using 4 classifiers for all the metrics, than using just one. The improvement is more evident in the TNR, i.e. in the female class recognition, presenting an advantage over the single classifiers. In any case for both databases the female class is worse represented in the training set, therefore, the resulting rates are unbalanced.

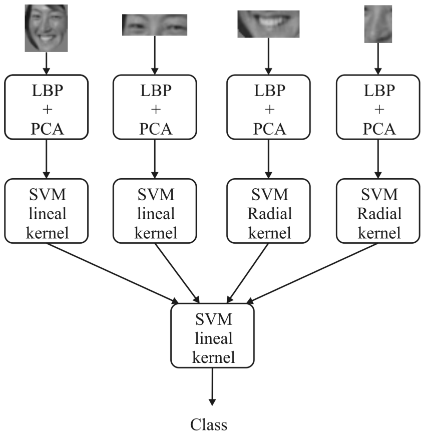


Fig. 4. Stacking based two classifiers combination.

Table 5. Classification rates per class and experiment using linear SVM for the LFW database

Fold	Accuracy	TPR	TNR	AUC
0	0.915	0.97	0.74	0.94
1	0.875	0.96	0.59	0.89
2	0.903	0.96	0.70	0.93
3	0.890	0.95	0.67	0.92
4	0.889	0.95	0.67	0.92
Mean	0.894	0.96	0.68	0.92

5 Conclusions

We have carried out a gender classification experiment on the LFW and MORPH datasets following the protocol defined by the BeFIT gender recognition challenge.

We have studied the performance using different areas of the normalized face, image resolutions, and face representations based on gray levels and LBP based preprocessing variants. The best performing single classifier makes use of the face pattern, performing slightly better than a classifier based on the eyes area.

Finally, we have designed a stacking based approach combining the single classifiers, achieving an accuracy of 88.6% and 92.4% respectively for the LFW and MORPH datbases. For the LFW database, those results do not outperform the studies presented in [20, 6]. The first study is restricted to frontal faces, and the second makes use of larger images and more time consuming representations such as Gabor jets. The results achieved for the MORPH database outperformed [5].

Table 6. Classification rates per class and experiment using linear SVM for the MORPH database.

Gray values				
Fold	**Accuracy**	**TPR**	**TNR**	**AUC**
0	0.921	0.98	0.62	0.94
1	0.921	0.98	0.62	0.94
2	0.929	0.98	0.66	0.95
3	0.922	0.98	0.62	0.95
4	0.927	0.98	0.64	0.95
Mean	0.924	0.98	0.64	0.95

LBP				
Fold	**Accuracy**	**TPR**	**TNR**	**AUC**
0	0.92	0.97	0.64	0.94
1	0.919	0.97	0.65	0.95
2	0.923	0.97	0.67	0.94
3	0.923	0.97	0.65	0.94
4	0.925	0.98	0.64	0.94
Mean	0.922	0.97	0.65	0.94

For future work, we expect to include cross-database tests (hopefully in the final version of the paper). We would also analyze more deeply the significance of the different face areas, and include histogram based face representation.

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. 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.
3. Modesto Castrillón, Oscar Déniz, Daniel Hernández, and Javier Lorenzo. A comparison of face and facial feature detectors based on the Viola-Jones general object detection framework. *Machine Vision and Applications*, 22(3):481–494, 2011.
4. Jie Chen, Shiguang Shan, Chu He, Guoying Zhao, M Pietikäinen, Xilin Chen, and Wen Gao. Wld: A robust local image descriptor. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1705–1720, September 2010.
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. 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.
7. Hazim Kemal Ekenel. Benchmarking facial image analysis technologies. http://face.cs.kit.edu/befit/, 2011.
8. A. Gallagher and T. Chen. Understanding images of groups of people. In *Proc. CVPR*, 2009.
9. F. Gosselin and P. G. Schyns. Bubbles: a technique to reveal the use of information in recognition tasks. *Vision Research*, pages 2261–2271, 2001.
10. Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The weka data mining software: An update. *SIGKDD Explorations*, 11(1), 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, University of Massachusetts, Amherst, October 2007.
12. Karl Ricanek Jr 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.
13. Y. Kirby and L. Sirovich. Application of the Karhunen-Loéve procedure for the characterization of human faces. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 12(1), July 1990.
14. Rainer Lienhart, Alexander Kuranov, and Vadim Pisarevsky. Empirical analysis of detection cascades of boosted classifiers for rapid object detection. In *DAGM'03, 25th Pattern Recognition Symposium*, pages 297–304, Magdeburg, Germany, September 2003.
15. 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.
16. 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.
17. Baback Moghaddam and Ming-Hsuan Yang. Learning gender with support faces. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(5):707–711, 2002. `http://ieeexplore.ieee.org/xpls/abs_all.jsp?tp=&arnumber=1000244&isnumber=21601`.
18. 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.
19. P. J. Phillips, H. Wechsler, J. Huang, and P. J. Rauss. The FERET database and evaluation procedure for facerecognition algorithms. *Image and Vision Computing*, 16(5):295–306, 1998.
20. Caifeng Shan. Learning local binary patterns for gender classification on real-world face images. *Pattern Recognition Letters*, (accepted), 2011.
21. Caifeng Shan, Shaogang Gong, and Peter W. McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and Vision Computing*, 27(6):803–816, May 2009.
22. Qian Tao and Raymond Veldhuis. Illumination normalization based on simplified local binary patterns for a face verification system. In *Proc. of the Biometrics Symposium*, pages 1–6, 2007.
23. M. Turk and A. Pentland. Eigenfaces for recognition. *J. Cognitive Neuroscience*, 3(1):71–86, 1991.
24. Paul Viola and Michael J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):151–173, May 2004.
25. Lior Wolf, Tal Hassner, and Yaniv Taigman. Descriptor based methods on the wild. In *In: Faces in Real-Life Images Workshop in ECCV*, 2008.
26. Zhiguang Yang and Haizhou Ai. *ADVANCES IN BIOMETRICS*, volume 4642, chapter Demographic Classification with Local Binary Patterns, pages 464–473. Springer, 2007.