Single- and Cross- Database Benchmarks for Gender Classification Under Unconstrained Settings

Pablo Dago-Casas, Daniel González-Jiménez, Long Long Yu
Multimodal Information Area
GRADIANT
{pdago,dgonzalez,longyu}@gradiant.org

José Luis Alba-Castro
Universidade de Vigo
jalba@gts.uvigo.es

Abstract

Gender classification is one of the most important tasks in automated face analysis, and has attracted the interest of researchers for years. Up to now, most gender classification approaches have been tested using single-database experiments, and on quite controlled datasets such as the FERET database, which are not representative of real world settings. However, a recent trend towards more realistic benchmarks has emerged within the face analysis community, leading to the appearance of databases and protocols such as the Labeled Faces in the Wild (LFW) database, and the so-called Gallagher's database, which comprises images collected from Flickr.

Contrary to LFW, where a standard protocol for gender classification has been established as one of the BeFIT challenges, there is no standard protocol in Gallagher's dataset, and a key contribution of this paper is to propose a standard 5-fold cross validation protocol for this database. Moreover, we provide cross-database experiments between Gallagher and LFW, as a way of assessing the performance of proposed algorithms in realistic conditions.

In addition, we revisit and compare appearance-based (pixels) and feature-based (Gabor and LBPs) descriptors combined with linear SVM-based and LDA-based classification, carrying out single-database (LFW and Gallagher's) and cross-database (Gallagher's $\rightarrow$ LFW and LFW $\rightarrow$ Gallagher's) experiments using the existing BeFIT challenge and the proposed dataset and protocols.

1. Introduction

Face-based gender classification has attracted the interest from the pattern recognition, machine learning and computer vision communities for years, due to the large number of areas where it is of potential interest: adaptive and dynamic advertising, biometrics, automatic indexing of multimedia content, and personalized HCI among others.

As stated in [4], automatic gender classification techniques can be divided into appearance-based and feature-based approaches. In the first case, the whole face patch (appropriately warped and normalized) is used as the feature for classification [4, 10, 17], while feature-based approaches exploit powerful image processing tools to extract facial features from the image, such as Haar-like [15], Gabor [11, 14, 21, 25, 26] or Local Binary Patterns (LBPs) [16, 20–22, 27] features.

The first approach to automatic gender classification from face images was SEXNET [10], where Golomb et al. used pixel information extracted from manually aligned 30x30 face images, and trained a Neural Network (NN) for classification. Using pixel information, but with automatically aligned faces, Moghaddam et al. [17] tested different classifiers on the FERET database, including Support Vector Machines (SVM) with Radial Basis Function (RBF) kernel, RBF Networks and linear classifiers. The SVM-RBF approach obtained the best classification rates for both high and low resolution images. Although SVMs have demonstrated good performance in gender classification, computationally cheaper classifiers such as Adaboost-based [3] or Linear Discriminant Analysis (LDA)-based [4] classification engines have proved to provide comparable accuracy to that of SVMs, while being significantly faster.

Regarding feature-based approaches, Gabor filters and LBPs have received considerable attention in the literature. In the context of Gabor-based gender classification, some authors convolve the whole image with a bank of filters [11, 21, 26], while others only extract features at sparse grid positions [14, 25]. Regarding LBPs, as stated above, an extense literature already exists [16, 20–22, 27]. Sun et al. [22] used LBPs with Self Organizing Maps (SOM) and Adaboost classifiers. [16] used LBPs in combination with SVM classifiers on automatically detected and aligned faces, while Yang et al. [27] used LBPs and Adaboost, built upon Chi square distance-based weak classifiers. Recently, [21] compared the performance of LBPs and Gabor features for age estimation and gender classification on a difficult dataset [8], which will be described below.

1.1. Benchmarking gender classification

At this point, it is necessary to briefly discuss training and testing databases and protocols, since it is a critical issue when benchmarking any face analysis technology.

Up to now, most approches for gender classification [3, 16, 17] have been tested on quite controlled datasets such as the FERET database [19]. However, these datasets are not representative of real conditions, and therefore it can not be expected that automatic systems trained and/or tested on such databases can generalize well when moving to real world images.

Therefore, a new trend has emerged in the face analysis community, focusing on the development of datasets and standard protocols for benchmarking face processing systems under realistic conditions: The Labeled Faces in the Wild (LFW) [12] for face recognition, Gallagher's DB for demographics estimation [8], the Face Detection Database (FDDB) for face detection, and the Dynamic Facial Expressions in the Wild for automatic expression analysis are just some examples of this new tendency. Moreover, initiatives such as the one taken by the Facial Image Processing and Analysis group (FIPA, `http://fipa.cs.kit.edu/`) in adapting existing datasets to new challenges (e.g. using the LFW dataset for benchmarking gender classification) are fostering fair competition in the research community.

The database described in [8] is very interesting since it comprises a large number of realistic images taken from Flickr (28231 people), and where every face has been manually labeled with its gender and age group. From now on, we will refer to this DB as Gallagher's database or Flickr database. Since acquisition conditions are very different among images, the database contains significant pose, illumination, expression, age and ethnic differences, making this dataset really challenging and representative of real world settings.

Within this trend towards real world testing, the database described in [8] and the adaptation of LFW's protocol provided in [1] constitute an appropriate framework for benchmarking gender classification in unconstrained conditions. Several works have already been using these datasets: apart from describing the database, [8] proposed a combination of appearance and context information for gender classification, achieving a correct classification rate of $\approx 74\%$. The tests carried out by Shan in [21] with LBP and Gabor features showed better performance than the method of [8]. Shan also tested LBP-based systems on LFW [20], achieving a correct recognition rate of 94.44%.

Apart from the importance of testing systems on realistic data, researchers have pointed out the usefulness of cross-database tests, in order to increase the significance and validity of the obtained results. In this direction, [4] performed a set of experiments using different databases (including FERET and other less constrained datasets), demonstrating that single database experiments are optimistically biased due to similar demography and capture conditions within a given database, and showing that performance degrades when cross-database experiments are carried out.

To the best of our knowledge, and contrary to LFW, there is no standard protocol in Gallagher's dataset and this provokes unfair comparisons between approaches. For instance, [21] takes into account every face whose interocular distance exceeds 24 pixels, but there is no public protocol specifying which images are used for training and which for testing, neither a n-fold cross validation scheme. This paper proposes a standard protocol for Gallagher's database following the 5-fold cross validation benchmark described in [1]. Moreover, we propose cross-database protocols using both datasets (Gallagher's $\rightarrow$ LFW and LFW $\rightarrow$ Gallagher's) for gaining significance and validity in the obtained results.

Furthermore, this paper revisits both appearance-based (pixels) and feature-based (Gabor and LBPs) descriptors combined with linear SVMs and LDA for gender classification, carrying out single-database and cross-database experiments using the existing BeFIT challenge and the proposed datasets and protocols.

This work is organized as follows. Section 2 details the system blocks of the gender recognition algorithms used for the experiments in this paper. Section 3 explains the benchmark databases and protocols proposed, while the specific experimental setup is exposed in section 4. Section 5 concludes the paper.

2. System building blocks

The gender classifier designed and tested in this work is divided in the following steps:

1. Face detection, alignment and normalization with two Region of Interest (ROI) sizes.
2. Feature extraction with three different descriptors.
3. Dimensionality reduction.
4. Classification with two methods.

2.1. Face detection and alignment

Faces were detected using the Viola-Jones (V&J) [24] implementation from OpenCV library [2]. After face detection, the eyes were automatically located, and the image was rotated and scaled so that the eyes lie in the same horizontal line, and the interocular distance is set to 45 pixels. After aligning and scaling, the face region is cropped to two different sizes, in order to test the influence of internal and external face features (in this work, external face features refer to face and chin contour, but do not include hair as is usual) in gender recognition. The smallest is 105×90, with eyes in coordinates $(22, 27)$ and $(57, 27)$. The other size we used is 120×105, with eyes in coordinates $(30, 35)$ and $(75, 35)$. Example images for the two used ROI sizes can be seen in Figure 1.

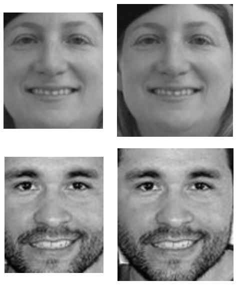

Figure 1. Example images for the two used ROIs. From left to right, 105×90 and 120×105 images.

2.2. Feature extraction

After face detection and alignment, three image descriptors are used for feature extraction: pixels, Gabor jets and LBPs, all of them used on gray level images with equalized histogram.

Pixels are commonly used as a baseline for new approaches, and have shown acceptable performance for gender classification combined with different classifiers.

As seen in Section 1, Gabor filters [7] have been used for automatic gender recognition using the whole filtered image or only a sparse grid of points. In this work we used a 10×10 uniform grid. At each point of this grid we obtained the modulus of the output of 40 complex Gabor filters (so called *jet*), using 5 frequencies and 8 orientations. The dimension of the whole feature vector, composed by the jets of the 10×10 grid, is 4000.

The third image descriptor tested in this work is LBPs [18]. To calculate the LBPs in this work we used LBP binary values from a neighboring region of 3x3 pixels, and obtained the 59-bins (one for non-uniform and 58 for uniform patterns) normalized histogram in 15×15 non-overlapping pixel blocks of the face image. Histograms are then concatenated, and the feature vector has dimension 2478 for 105×90 images and 3304 for 120×105 images.

After feature extraction, Principal component Analysis (PCA) [13] is applied for dimensionality reduction, as in [4]. The selection of the optimal percentage of energy kept in this stage will be explained in Section 3.2

2.3. Classification

Finally, after PCA we tested two different classification methods:

- Linear Support Vector Machine (SVM) [5],
- Linear Discriminant Analysis (LDA) [6].

SVMs have been used in most recent works regarding gender classification [4, 15, 21], achieving good classification rates even in unconstrained conditions [21]. RBF SVMs are used more often than linear SVMs, but we decided to use the latter because RBF SVMs need one parameter more to be selected (parameter γ), and the difference between them does not justify, for these experiments, the increase in the computational costs. The other tested classifier is LDA, a dimensionality reduction technique that provides a linear projection of the data into a subspace maximizing the inter-class separability while minimizing the intra-class dispersion. LDA classifier is faster to train than linear SVM, and has also shown good performance for gender classification in different conditions [4].

3. Benchmark

In this section we revisit the databases used for the experiments, we propose a test protocol and explain the metrics used to evaluate the results.

3.1. Used databases

For our experiments we used Gallagher's [8] and Labeled Faces in the Wild (LFW) [12] databases, both containing images taken in unconstrained conditions, as can be seen in the examples shown in Figure 2. Gallagher's database is composed by 28231 labeled faces collected from Flickr images, and is publicly available. It has been previously used for gender classification in unconstrained conditions [8, 21], but there is not a common and defined protocol for the experiments, and the results obtained by different authors using this database cannot be fairly compared. LFW contains 13233 labeled images from 5749 individuals collected from the web. There is prior work in gender recognition using this database [20], and a benchmarking protocol to standardize gender recognition experiments in this database is proposed in [1].

Figure 2. Examples faces from Gallagher's (upper row) and LFW (bottom row) databases.

Since there is not a standard protocol for gender classification in Gallagher's database, in this work, and following [1], we propose a 5-fold division over a subset of the original dataset, obtained as follows. We used the automatic face detection and alignment method explained in Section 2.1 on Gallagher's labeled faces, and removed those faces whose interocular distance in the original image was less than 20 pixels (to eliminate low resolution faces). This resulted in an image set composed by 15579 faces. To have equal number of male and female faces we randomly removed some of the male faces, resulting this in a final dataset containing 14760 images. We divided this image collection into 5 equally sized folds, each one comprising 2952 images. For this purpose, we randomly selected the images in each fold, keeping a balanced gender distribution (1476 male and 1476 female faces in each fold). The information about the fold distribution is available upon request to the research community. This information allows other authors to evaluate their own procedure using the proposed folds.

For the LFW database we used the 5 folds proposed in [1], and applied the same detection and alignment algorithm explained in Section 2.1. With the used parameters, some of the faces were not detected by V&J detector, and this results in a dataset composed by 13088 images, 10129 male and 2959 female. As it can be seen, there is a clear imbalance between genders.

3.2. Test protocol

As said before, we carried out two kinds of tests: single- and cross-database tests. In the following, we describe both single- and cross-database protocols, besides exposing the metrics used to evaluate the results.

3.2.1 Single-database tests

In single-database tests (both for Gallagher and LFW databases), we used 4 folds for training and 1 for testing. In the training stage, in order to select the optimum values of the parameter C of the SVM (when using SVMs classifier) and the percentage of energy kept in PCA stage, we used the following validation procedure. One of the training folds was selected as validation fold and, using the three remaining folds, we trained a classifier for each value (or pair of values in the SVM case) of the parameters we wanted to test and tested them in the validation fold. This was repeated using the four training folds as validation fold and results were averaged. Then, those values of the parameters that obtained the best average classification accuracy were selected.

After validation, the optimum parameters were used to train a classifier using the whole training set (4 folds), and its performance was tested in the fifth fold. This whole procedure (depicted in Figure 3) was repeated for the 5 possible combinations of training and test folds, and the results were averaged.

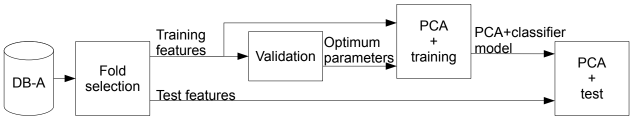

Figure 3. Block diagram of the procedure of single-database tests. This procedure is repeated using every fold as test fold.

3.2.2 Cross-database tests

In cross-database tests we used the same fold division as in single-database tests, and applied an equivalent procedure to fairly compare the results. We used 4 folds (from the same database) for training a classifier, using the optimum parameters obtained in the single-database experiments. Then the classifier was tested using all the folds of the other database. This procedure (depicted in Figure 4) was repeated for the 5 possible 4-folds combinations from the training database, and results were averaged. Then, the training database was used as test database and vice versa.

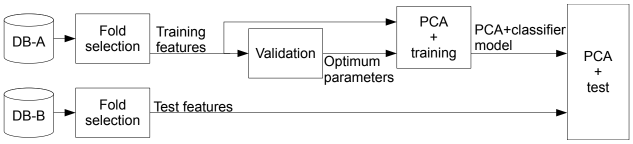

Figure 4. Block diagram of the procedure of cross-database tests. This procedure is repeated using every fold from database B as test fold.

3.2.3 Metrics

To compare the performance of the different tested schemes, we used the following metrics [9]: Accuracy (ACC), True Positive Rate (TPR) and True Negative Rate (TNR).

$$ACC = \frac{TP + TN}{P + N}, \tag{1}$$

$$TPR = \frac{TP}{P}, \tag{2}$$

$$TNR = \frac{TN}{N}, \tag{3}$$

being TP the number of test samples correctly classified as positive, TN the number of test samples correctly classified as negative, P the total number of positive test samples and N the total number of negative test samples.

4. Experiments

In this section we explain the experiments carried out following the protocol exposed in Section 3. First, we comment the single database tests made in Gallagher's and LFW databases and the obtained results. Then, we expose cross-database tests and comment their results.

4.1. Single-database experiments

The objective of this experiments is to determine the performance of the different configurations proposed in this work for:

- Image descriptors: pixels, Gabor jets and LBPs.
- Cropped face size: 105×90 and 120×105.
- Classifier: SVM and LDA.

4.1.1 Gallagher's database

The results of the tests for Gallagher's database are shown in Table 1. Results show that Gabor jets and LBPs perform better than pixels, and also how using 120×105 ROIs improves classification accuracy (this difference is more relevant for LBPs than for the other descriptors). TPR and TNR are balanced, which means the classifier does not favor one of the classes against the other. Results also show that for this database, PCA+SVM and PCA+LDA schemes have similar performance.

We obtained a classification accuracy of 79.16% for pixels, while Gallagher and Chen [8] obtained 69.96% using only appearance, and 74.1% combining appearance and context information in Gallagher's database. For Gabor jets and LBPs we achieved 86.61% and 86.34% respectively, while Shan [21] obtained 75.7% for boosted Gabor features and 77.4% for boosted LBPs, both combined with RBF SVMs. Nevertheless, these results must be compared carefully, because they have not been obtained in a common framework and, therefore we are not able to make fair comparisons.

In fact, Gallagher and Chen used 3500 faces for training, randomly selected and having an uniform distribution among the different age groups. For test, they selected an independent set of 1050 images. While Gallagher and Chen did not take into account the resolution of the faces, Shan [21] selected those images from Gallagher's database with interocular distance larger than 24 pixels, and selected a training set containing 9336 faces and a test set with 2744 faces, having a balanced gender distribution.

Faces in Gallagher's database have also age labels, and we find relevant to compare results by ages. Each individual is included in one of these age groups: 0-2, 3-7, 8-12, 13-19, 20-36, 37-65 or 65+ (in years). Due to lack of space, in Figure 5 we only show the accuracy for each age group of the classifier using LBPs, 120×105 images and PCA+SVM scheme, but the other tested schemes show similar behavior. Accuracy in middle-aged adults groups (20-36 and 37-65) rises to 90%, while the lowest classification rates are located in 0-2, 3-7 and 8-12 groups (between 60 and 70% of accuracy). The low number of images in children's groups (see Figure 6) could explain this fact, and although groups 13-19 and 65+ have also few images, their faces are more similar to those from middle-aged adults, and gender recognition rates are not so influenced by the number of images.

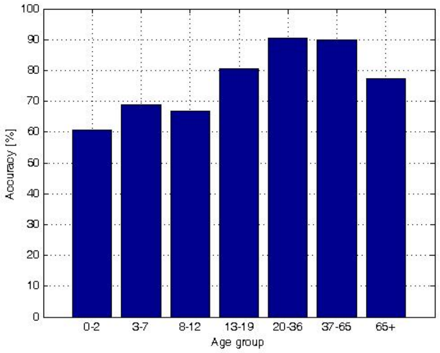

Figure 5. Accuracy (%) for the different age groups using LBPs, 120×105 images and PCA+SVM scheme training and testing in Gallagher's database.

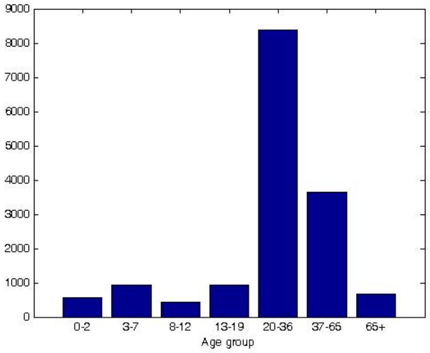

Figure 6. Histogram of the number of images in each age group of Gallagher's database.

4.1.2 LFW database

Table 2 shows the obtained results using the LFW database. As happened in the experiments with Gallagher's database, Gabor jets and LBPs perform better than pixels, and 120×105 images offer higher classification rates (94.01% for Gabor jets and 93.83% for LBPs), similar to those obtained by Shan in [20] (94.44% using LBP-based systems). Although accuracy is higher than in Gallagher's tests (the difference is not so significant if we only take into account the results middle-aged adults in Gallagher's database), the imbalance between positive (male) and negative (female) samples in LFW makes the classifier favor the positive class against the negative one, as show the TPR and TNR values.

There are several options to deal with imbalanced datasets [23]. Undersampling the most populated class is one of them, but this means discarding useful information and reducing the number of images in the database. We decided to use a weighted SVM (W-SVM) for classifying, although selecting the optimum weights for the two classes is not a trivial issue [23], we decided to assign the negative class three times the weight of the positive one (the same ratio as the number of samples from each class) to test its effectiveness, but it would be necessary to do an exhaustive search to obtain the optimum weights. As can be seen in Table 3, with the proposed scheme the accuracy is a bit lower than using the non-weighted SVM, but TPR and TNR are now more balanced (the difference between them is around 5%, instead of the 15% obtained in the previous tests).

4.2. Cross-database experiments

In order to check the generalization properties of the trained systems in realistic conditions, we run cross-database tests using Gallagher's for training and LFW for testing and vice versa. For this purpose, we used the procedure explained in Section 3.2. By using the same images for training in both single- and cross-database tests, we are able to check the generalization properties of the proposed schemes for each database without the bias introduced by the number of images used in the training stage.

Table 4 shows the obtained results when training with Gallagher's and testing with LFW. As in the previous experiments, Gabor jets, LBPs and 120×105 ROI perform better. Accuracy (*e.g.* 89.77% for LBPs) is higher than in intra Gallagher's tests, and similar to the accuracy obtained for middle-aged adults (around 90%, see Figure 5). Nevertheless, TPR and TNR are now very unbalanced, influenced by the different number of positive (male) and negative (female) samples in the test database (LFW). Again, differences among PCA+SVM and PCA+LDA schemes are not significant.

The obtained results when training in the LFW and testing in Gallagher's database are shown in Table 5. Results have worsen significantly in comparison with intra tests in LFW, but the behavior is similar, obtaining Gabor jets and LBPs higher accuracy than pixels, and having slightly better performance with 120×105 ROIs. As seen before, LFW database has much more positive than negative samples, and this clearly reflected in TPR and TNR values for PCA+SVM classifier. Although intra-LFW tests results using W-SVM were balanced, results for cross-database tests show how TNR is now higher than TPR, probably because the classifier is slightly overfitted, and is not able to generalize when testing in a different database. When using PCA+LDA scheme, the imbalance among TPR and TNR is not so acute, probably due to the minor complexity of LDA compared to SVM (less parameters were trained with an unbalanced dataset).

Analyzing results by age group (see Figure 7), the lack of children images in LFW database (used for training) makes accuracies of groups 0-2, 3-7 and 8-12 drop significantly (accuracy is below 60%), and is also low for 65+ subjects (around 75%). For midle aged-adults, the classifier perform well, obtaining good classificaton rates.

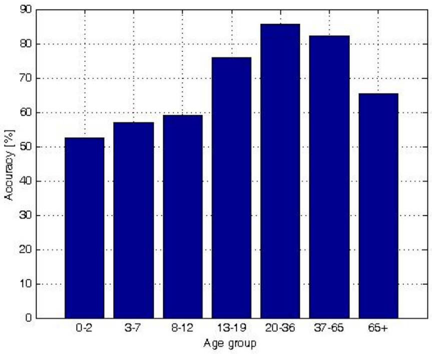

Figure 7. Accuracy (%) for the different age groups using LBPs, 120×105 images and PCA+SVM scheme training with LFW and testing in Gallagher's database.

5. Conclusions

In this work we have presented a benchmarking protocol for Gallagher's dataset following the one proposed in [1] for LFW database. We have shown the good generalization properties of the proposed protocol with this database, allowing to obtain accurate estimations of the performance of the tested schemes in real application frameworks.

All tests have shown a similar behavior for the different descriptors. Gabor jets and LBPs obtain similar accuracies, and perform better than pixels. In addition, using 120×105 ROI slightly improves the results compared to those obtained using 105×90 ROIs, specially for LBPs. In general, PCA+SVM works better than PCA+LDA scheme, but there is not a significant difference, except in intra LFW tests where PCA+LDA obtains more balanced values of TPR and TNR.

We have shown how the proposed protocol for Gallagher's database does not favor one of the classes against the other, due to the balanced gender distribution in each of the proposed folds and in the whole database. The imbalance in the LFW database is an added issue to deal with when working with it.

Cross-database tests have shown how results obtained in intra Gallagher's experiments with the proposed protocol offer a good estimation of its performance in real application frameworks (the estimation is even more accurate if we only take into account age groups present in LFW database). In addition, cross-database tests emphasized the effects of the imbalance in the LFW dataset, causing a bias of TPR and TNR when using it for training or test. Even training LFW with W-SVM, that improves balanced performance in the intra-LFW tests, seems to slightly overfit the system to the training set because it introduces imbalance in the opposite direction when testing in Gallaghers database and improves only slightly for LBP. This behavior highlights the difficulty of dealing with an imbalanced dataset as LFW for generalizing gender classification results.

For these reasons, we propose the use of Gallaghers database for gender classification in unconstrained conditions. This database has a balanced number of male and female samples, besides a great demographic, pose and conditions variability.

Although the images using 105×90 and 120×105 ROI are not very different, the slight improvement of the classification whit 120×105 ROI containing face contour information can motivate future work combining appearance and shape information. The use of larger ROIs can also be tested. Regarding to the classifier, boosting techniques as those applied in [21] should be tested in single- and cross-database experiments, in order to determine its applicability.

Acknowledgements

This work was partially funded by Xunta de Galicia under projects VISAGE (10TIC008CT) and CN 2011/019, and by the Spanish Ministry of Science and Innovation under project AREA (TEC2008-05894).

References

[1] http://fipa.cs.kit.edu/downloads/LFW-gender-folds.dat.

[2] http://opencv.willowgarage.com/wiki.

[3] S. Baluja and H. Rowley. Boosting Sex Identification Performance. *International Journal of Computer Vision*, 71:111–119, January 2007.

[4] J. Bekios-Calfa, J. Buenaposada, and L. Baumela. Revisiting Linear Discriminant Techniques in Gender Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33:858–864, 2011.

[5] C. Cortes and V. Vapnik. Support Vector Networks. *Machine Learning*, 20(3):273–297, 1995.

[6] R. Duda and P. Hart. *Pattern Classification and Scene Analysis*. John Wiley & Sons, 1973.

[7] D. Gabor. Theory of Communication. *Journal of Institute for Electrical Engineering*, 93, 1946.

[8] A. Gallagher and T. Chen. Understanding Images of Groups of People. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[9] T. Gehrig, M. Steiner, and H. Ekenel. Draft: Evaluation Guidelines for Gender Classification and Age Estimation. http://fipa.cs.kit.edu/downloads/befit-evaluation_guidelines.pdf.

[10] B. A. Golomb, D. T. Lawrence, and T. J. Sejnowski. SEXNET: A Neural Network Identifies Sex from Human Faces. In *Proceedings of the 1990 conference on Advances in neural information processing systems 3*, pages 572–577, 1991.

[11] G. Guo, C. Dyer, Y. Fu, and T. Huang. Is Gender Recognition Affected by Age? In *Proceedings of IEEE 12th International Conference on Computer Vision Workshops*, pages 2032–2039, 2009.

[12] G. B. Huang, T. B. M. Ramesh, and E. Learned-Miller. Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. Technical report, University of Massachusetts, Amherst, 2007.

[13] I. Jolliffe. *Principal Component Analysis*. Springer, 2002.

[14] M. Lyons, J. Budynek, A. Plante, and S. Akamatsu. Classifying Facial Attributes Using a 2-D Gabor Wavelet representation and Discriminant Analysis. In *Proceedings of International Conference on Automatic Face and Gesture Recognition*, pages 202–207, 2000.

[15] E. Mäkinen and R. Raisamo. Evaluation of Gender Classification Methods with Automatically Detected and Aligned Faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(3):541–547, 2002.

[16] E. Mäkinen and R. Raisamo. An Experimental Comparison of Gender Classification Methods. *Pattern Recognition Letters*, 29:1544–1556, 2008.

[17] B. Moghaddam and M.-H. Yang. Learning Gender With Support Faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):707–711, 2002.

[18] T. Ojala, M. Pietikäinen, and T. Mäenpää. A Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002.

[19] P. Phillips, H. Moon, S. Rizvi, and P. Rauss. The FERET Evaluation Methodology for Face-Recognition Algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(10):1090–1104, Oct. 2000.

[20] C. Shan. Gender Classification on Real-Life Faces. In *Proceedings of Advanced Concepts for Intelligent Vision Systems*, volume 6475/2010, pages 323–331, 2010.

[21] C. Shan. Learning Local Features for Age Estimation on Real-Life Faces. In *Proceedings of the 1st ACM International Workshop on Multimodal Pervasive Video Analysis*, pages 23–28, 2010.

[22] N. Sun, W. Zheng, C. Sun, C. Zou, and L. Zhao. Gender Classification Based on Boosting Local Binary Pattern. In *Proceedings of 3rd International Symposium on Neural Networks*, volume 2, pages 194–201, 2006.

[23] Y. Tang, Y.-Q. Zhang, N. Chawla, and S.Krasser. SVMs Modelling for Highly Imbalanced Classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics - Special issue on human computing archive*, 39:281–288, 2009.

[24] P. Viola and M. Jones. Robust Real Time Detection. In *Second International Workshop on Statistical and Computational Theories of Vision-Modeling, Learning, Computing, and Sampling*, 2001.

[25] L. Wiskott, J.-M. Fellous, N. Krüger, and C. V. der Malsburg. Face Recognition and Gender Determination. In *Proceedings of First International Workshop on Face and Gesture Recognition*, pages 92–97, 1995.

[26] B. Xia, H. Sun, and B.-L. Lu. Multi-View Gender Classification Based on Local Gabor Binary Mapping Pattern and Support Vector Machines. In *IEEE International Joint Conference on Neural Networks*, pages 3388–3395, 2008.

[27] Z. Yang and H. Ai. Demographic Classification With Local Binary Patterns. In *Proceedings of International Conference on Biometrics (ICB)*, pages 464–473, 2007.

Table 1. ACC, TPR and TNR (%) in Gallagher's database for the three descriptors (pixels, Gabor jets and LBPs), the two image sizes (105 × 90 and 120 × 105) and the two classification methods (PCA+SVM and PCA+LDA).

		Pixels		Gabor jets		LBPs	
		105 × 90	120 × 105	105 × 90	120 × 105	105 × 90	120 × 105
PCA+SVM	ACC	79.06	79.16	85.58	86.61	84.55	86.34
	TPR	78.01	78.16	85.62	87.24	84.26	86.69
	TNR	80.11	80.16	85.54	85.98	84.84	85.99
PCA+LDA	ACC	78.74	79.09	85.33	86.58	84.16	86.02
	TPR	77.18	77.14	85.46	86.91	83.24	85.56
	TNR	80.30	81.04	85.19	86.25	85.08	86.49

Table 2. ACC, TPR and TNR (%) in LFW database for the three descriptors (pixels, Gabor jets and LBPs), the two ROI sizes (105 × 90 and 120 × 105) and the two classification methods (PCA+SVM and PCA+LDA).

		Pixels		Gabor jets		LBPs	
		105 × 90	120 × 105	105 × 90	120 × 105	105 × 90	120 × 105
PCA+SVM	ACC	89.17	89.24	93.47	94.01	92.91	93.83
	TPR	95.40	95.40	97.23	97.47	97.23	97.01
	TNR	67.86	68.13	80.60	82.16	78.14	82.97
PCA+LDA	ACC	86.96	86.47	92.98	93.41	90.93	92.80
	TPR	88.15	87.52	95.17	95.48	92.20	93.98
	TNR	82.87	82.87	85.47	86.35	86.58	88.75

Table 3. ACC, TPR and TNR (%) in LFW database for the three descriptors (pixels, Gabor jets and LBPs), the two ROI sizes (105 × 90 and 120 × 105) using the classification method PCA+W-SVM.

		Pixels		Gabor jets		LBPs	
		105 × 90	120 × 105	105 × 90	120 × 105	105 × 90	120 × 105
PCA+W-SVM	ACC	86.31	86.45	92.47	92.96	91.09	92.96
	TPR	87.20	87.38	93.55	94.10	92.05	93.84
	TNR	83.24	83.28	88.75	89.05	87.80	89.96

Table 4. ACC, TPR and TNR (%) for cross-database tests training in Gallagher's and testing in LFW. Results are shown for the three descriptors (pixels, Gabor jets and LBPs), the two ROI sizes (105 × 90 and 120 × 105) and the two classification methods (PCA+SVM and PCA+LDA).

		Pixels		Gabor jets		LBPs	
		105 × 90	120 × 105	105 × 90	120 × 105	105 × 90	120 × 105
PCA+SVM	ACC	81.40	80.70	88.27	89.64	86.74	89.77
	TPR	82.11	81.43	90.10	91.71	88.59	91.80
	TNR	78.98	78.18	82.02	82.56	80.41	82.85
PCA+LDA	ACC	81.07	80.47	87.91	89.27	85.87	89.15
	TPR	81.93	81.41	90.12	91.50	87.69	91.13
	TNR	78.14	77.22	80.34	81.64	79.64	82.37

Table 5. ACC, TPR and TNR (%) for cross-database tests training in LFW and testing in Gallagher's. Results are shown for the three descriptors (pixels, Gabor jets and LBPs), the two ROI sizes (105 × 90 and 120 × 105) and the three classification methods (PCA+SVM, PCA+LDA and PCA+W-SVM).

		Pixels		Gabor jets		LBPs	
		105 × 90	120 × 105	105 × 90	120 × 105	105 × 90	120 × 105
PCA+SVM	ACC	71.52	72.09	79.08	80.15	77.21	79.65
	TPR	85.50	84.39	83.46	84.22	85.39	85.82
	TNR	57.54	59.79	74.69	76.08	69.04	73.48
PCA+W-SVM	ACC	72.78	72.97	78.44	79.66	78.13	80.17
	TPR	67.61	66.20	72.20	73.74	73.54	76.87
	TNR	77.94	79.74	84.67	85.59	82.71	83.47
PCA+LDA	ACC	71.87	72.56	79.52	80.59	78.68	81.02
	TPR	65.66	66.16	78.09	78.80	76.22	80.68
	TNR	78.08	78.96	80.96	82.38	81.13	81.35