Within- and Cross- Database Evaluations for Gender Classification via BeFIT Protocols

Nesli Erdogmus, Matthias Vanoni, Sébastien Marcel,
Idiap Research Institute
Centre du Parc, Rue Marconi 19, CH-1920, Martigny, Switzerland
Email: nesli.erdogmus, matthias.vanoni, marcel@idiap.ch

***Abstract*—With its wide range of applicability, gender classification is an important task in face image analysis and it has drawn a great interest from the pattern recognition community. In this paper, we aim to deal with this problem using Local Binary Pattern Histogram Sequences as feature vectors in general. Differently from what has been done in similar studies, the algorithm parameters used in cropping and feature extraction steps are selected after an extensive grid search using BANCA and MOBIO databases. The final system which is evaluated on FERET, MORPH-II and LFW with gender balanced and imbalanced training sets is shown to achieve commensurate and better results compared to other state-of-the-art performances on those databases. The system is additionally tested for cross-database training in order to assess its accuracy in real world conditions. For both within- and cross-database experiments on LFW and MORPH-II, the BeFIT protocols are utilized.**

I. Introduction

While human faces carry enough information to allow gender perception, as we experiment successfully on a daily basis, automated gender classification based on face images is still an open research problem which addresses a diverse range of applications. For instance, it could serve as a pre-processing stage in a face recognition system to prune a large biometric database in order to reduce the search load. In marketing, it can help collecting demographic statistics or adapting the content of an advertisement in in real-time according to the audience. In human-robot interaction, it can enable robots to adopt a suitable behavior depending on the user's gender.

The gender classification tasks in pattern recognition, is a binary task and consists of labelling images of human faces as "male" or "female". This task can be challenging because the face images are subject to a wide range of variations in terms of image quality, pose, illumination and expression differences and presence of occlusions such as make-up or facial hair. Moreover, it is also influenced by age and ethnicity [10]. Hence, gender recognition algorithms may suffer from data dependency with a risk of poor generalization. Face databases, in that sense, are of crucial importance. In [8], a benchmarking and evaluation protocol for gender classification is proposed in the context of BeFIT benchmarks. In this study we follow these protocols in order to evaluate the proposed system. Additionally, considering that fair comparison of methods and reproducibility of the results benefit to both the readers and the authors, the source code for these tests is made available.

Our contribution in this paper is two folds. Firstly, a large number of parameter combinations for cropping the face image and extracting the Local Binary Pattern Histogram Sequences (LBPHS) are tested to find which configuration is the most suitable for gender classification task. For this purpose, an in-depth analysis is carried out on BANCA [3] and MOBIO [14] databases. Secondly, the final system constructed with the obtained optimal parameters is tested with within- and cross-database experiments using MORPH-II [19], LFW [11] and FERET [18] databases.

The rest of the paper is organized as follows: In Section II details on the related work is given. In Section III, the proposed method is presented. Experimental results for within- and cross- database evaluation are provided in Section IV. Finally, in Section V the paper is concluded with remarks on the obtained results and the future work.

II. Related work

For the sake of simplicity, the related work on gender classification is examined with respect to 3 main sub-tasks: feature extraction, feature selection and classification. The feature extraction step turns images into feature vectors and the feature selection step reduces their dimensionality. Finally, the classifier determines a boundary in the reduced feature space so that the images are labelled as one of the two classes.

A. Feature extraction

The simplest feature to use in image analysis is raw pixel intensities. In this method, the size of the cropped face has a direct impact on the resulting feature vector's dimensionality, so the use of intensity as a feature is usually observed on small images. To our knowledge, Golomb et al. [9] published the first study on gender classification which uses face images of 30×30 pixels as features and a Neural Network (NN) as a classifier. They obtained 91.9% accuracy compared to 88.4% accuracy achieved by humans. Later, Moghaddam et al. [15] were able to reach 96.0% of accuracy on even smaller images of 21×12 pixels, using a Support Vector Machine (SVM).

Intensity values might be the most appropriate features, in case gender recognition is required on very low resolution images. However, several comparative studies [13], [28], [6] prove that even better features exist such as Gabor, LBP or their combination.

The use of Gabor features in image processing has been introduced by Daugman [7] and the effectiveness of such low level features has been proved for many pattern recognition tasks such as iris, fingerprint or face recognition [21].

In 1996, Wiskott et al. [26] proposed to apply Gabor features for gender classification and reported 90.2% accuracy

on FERET. More recently in [10], a variant of Gabor features is used (Biological Inspired Features (BIF) [20]) and 98.3% accuracy is achieved on MORPH-II using 60×60 images.

LBP codes are texture descriptors introduced in [16] and further improved in [17] by Ojala et al.. LBP operator converts the intensity value of a pixel into a binary code based on the difference of intensities with respect to its neighbors. Furthermore, these binary codes can be grouped in a histogram that compactly describes the image content. In order to keep spatial information, the image can also be divided into sub-regions and the LBP histogram of each region can be concatenated, resulting in a LBPHS vector. This method has been widely used for gender classification [23], [22], [28].

In [22], the eyebrows, the eyes, and the region between the nose and the mouth are claimed to be the most discriminative parts of the face for gender classification using LBP features. The best results reported so far with LBP features are [28] with 95.60% using FERET, CAS-PEAL and BCMI databases (controlled) and [22] with 94.81% on LFW (uncontrolled).

Interestingly, the combination of Gabor and LBP by extracting LBPHS from Gabor images leads to very discriminative features as demonstrated in [28], [27]. In [28] Gabor-based, LBP-based, and LGBP-based algorithms are compared on the FERET, CAS-PEAL and BCMI databases, and 92.17%, 95.60% and 99.84% overall accuracies are obtained for each feature type, respectively, which proves the supremacy of LGBP features in gender classification with controlled images.

B. Feature selection methods

Once the features are extracted, the next step is to reduce the dimensionality of the feature vector by selecting the more discriminating features and/or by projecting the features into a lower dimensionality space, which is expected to achieve better performance in terms of accuracy.

Some examples of feature selection methods applied in previous gender classification studies are Independent Component Analysis (ICA) on intensity values in [12], Canonical Component Analysis (CCA) on BIF in [10], Linear Discriminant Analysis (LDA) in [28], [27] and AdaBoost in [22], [13].

C. Classification

The selection of adapted features for the task of gender classification has a significant influence on the final performance of the system, and so has the choice of the classifier. In the literature, several classification methods have been tested for gender such as NN [9], Discriminative analysis of the Canonical Correlation (DCC) [4], CCA [10], Partial Least Squares (PLS) [10], AdaBoost [23] or Support Vector Machines (SVM) [22].

In particular, SVM have been proved to outperform most of the other classifiers. It is a supervised learning algorithm introduced by Cortes and Vapnik [5] in 1995. Given a labelled training set of data, it aims at finding a linear hyperplane that best separates them in the feature space. In case of a non-linear problem, then the SVM algorithm maps the data in a higher dimensional space and finds a linear separator in that space. A "kernel function" has to be chosen *a priori*, thus leading to different kinds of classifiers depending on the chosen kernel.

Moghaddam and Yang [15] were the first to use an SVM classifier for gender classification. They reported 96.6% of accuracy on the FERET database, on 21×12 face images. Since Moghaddam, SVMs gained in popularity and numerous types of features and SVMs have been tested for gender classification [22], [28].

III. Proposed approach

In general, the problem of gender classification from face images can be broken down into four stages: detection (localization) of the face, pre-processing of the facial region, feature extraction and classification (Figure 1).

The first step which aims to detect/locate a face in an image is left out of scope of this study. In all experiments, manually annotated eye positions are utilized.

As for the pre-processing step, only cropping and geometric normalization is considered. It means that faces are rotated and aligned in the images using the manually annotated eye positions. In this way, eyes are brought to the same location in each face image and the faces are brought to the same scale in upright position. This approach is widely employed in face studies and in [13] the authors prove the positive impact of face normalization on gender classification by conducting experiments on the FERET database.

On the other hand, to the best of our knowledge, there are no studies that extensively compare different cropping parameters such as image size, resolution and aspect ratio. In the previous studies, these are most commonly chosen *a priori* at the beginning of the experiments and no justification is presented [10], [28]. In this work, we aim to find the optimum configuration for cropping and geometric normalization for gender classification.

In the third step, discriminative and representative features are selected and extracted from the face images. As mentioned in Section II, LBP is a powerful local texture descriptor which has been widely employed in face image analysis, including gender classification studies [1], [22]. Following this trend, in our study, we analyze the effectiveness of LBPHS and their variations as feature vectors to be classified into genders.

Similar to pre-processing, extraction of LBPHS is also based on a handful of parameters, such as types of LBP, radius, number of sampling points and number of blocks that the images are segmented into. In most of the existing work that utilize LBPHS for gender classification, these parameters are again set without any justification [1], [28]. On the other hand, there are also several studies for which LBPHS are extracted using different configurations and most discriminative bins are selected via a boosting algorithm [22], [23].

Regular LBP assigns a code to each image pixel by comparing its intensity with those of its neighbors. The distance of the neighboring pixels to the center and the number of points sampled from this circle are controlled by the LBP radius (R) and the number of sampling points (P), respectively (Figure 2). *Modified* LBP operates exactly the same except that the comparison is done to the average intensity of neighboring points instead of the center pixel [25].

An LBP code is called *uniform* when it contains at most two bitwise transitions and *non-uniform* otherwise. In [17],

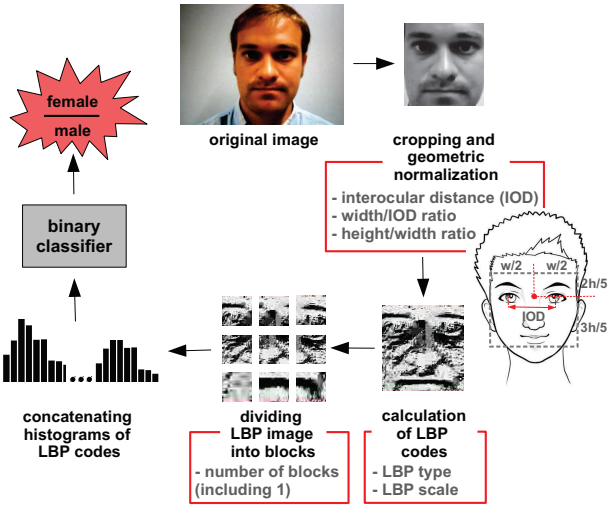

Fig. 1. The gender classification system is illustrated after broken down into its components. The optimized parameters from each step are given in red boxes.

Ojala et al. show that non-uniform patterns do not occur often enough to yield to reliable statistics. Furthermore in [22], the validity of uniform patterns is verified for gender classification. By adopting Adaboost to learn discriminative LBPHS bins, a strong classifier is built with $\text{LBP}_{8,2}$ operator which performs similarly to that of uniform $\text{LBP}_{8,2}$. Thus, in our experiments, we only consider uniform patterns and use a single label for all non-uniform ones. This results in N different labels (histogram bins), formulated by $N = P \times (P-1) + 2$ where P denotes the number of sampling points.

Once the LBP images are generated by replacing each pixel with its LBP code, the feature vectors can be computed in two different ways. In the first option, a single histogram is created using all pixels in the image. This results in a feature vector of length equal to the number of bins (N). In the second option, the image is divided into blocks and the final feature vector is formed by concatenation of histograms from each of the blocks. In this case, the length of the feature vector becomes $N \times M^2$ where M is the number of blocks in each axis, since same number of blocks are used in x and y directions. In order to simplify the process, the first option is regarded as a special case of the second option in which M is equal to 1.

In this study, we seek for the best configurations for cropping and LBP parameters together. Towards this end, we utilize two different databases and conduct a grid search for 6 parameters in total (Figure 1). They are listed as the following, for cropping and geometrical normalization:

- Inter-occular distance (IOD): [10, 20, 30]
- width-to-IOD ratio: [1.5, 2, 2.5, 3, 3.5]
- height-to-width ratio: [$\frac{1}{2}$, $\frac{1}{1.5}$, 1, 1.5, 2]

and for feature extraction:

- LBP type: [regular, modified]
- Number of sampling points and LBP radius (P,R): [(8,1), (8,2), (16,2)]
- Number of blocks in each axis (M): [1, 2, 4, 6, 8, 10, 12, 14, 16, 18, 20]

All parameters listed above and their corresponding options result in more than 4000 valid configurations with compatible values. Each of these configurations are tested on two different databases and ranked according to their performances in terms of average correct rate (please refer to Section IV-B). The optimum set of parameters is selected as the one with the best overall rank.

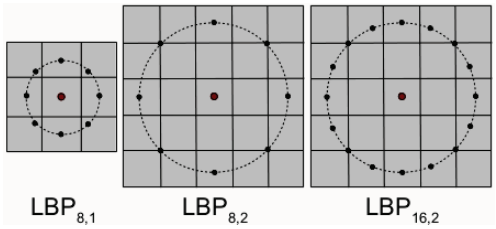

Fig. 2. Different LBP radiuses, R and number of sampling points, P ($\text{LBP}_{P,R}$)

Regarding the fourth and the last step of a gender classification system, a classifier is to be selected. In [24], SVM is compared to three different classifiers (a Bayes classifier, a NN classifier and a classifier based on LDA) and proved to give the best performance. Similarly in [15], Moghaddam et al. show that SVM is superior to various classifiers, such as nearest neighbor, Fisher linear discriminant and radial basis function (RBF) networks. In the same study, the authors additionally prove that SVM with RBF kernel works better than polynomial kernels of different degrees. In view of these results, SVM with RBF kernel is chosen to be utilized for our gender classification system.

IV. Experiments

Three sets of experiments (E1, E2 and E3)) are conducted:

- E1: In order to find optimal cropping and LBP parameters, firstly, a grid search is performed.
- E2: The findings of the first set of experiments are applied to build and test the final system using each database separately.
- E3: The final system is tested by cross-database experiments (training with one and evaluating with another)

All of the experiments are implemented using the free signal-processing and machine learning toolbox Bob[1] [2]. The source code for these experiments is available as one of its satellite packages[2] to foster replicable research.

A. Databases and protocols

The first set of experiments, E1, that aim to find ideal algorithm parameters are conducted on the images extracted from the video corpus of BANCA [3] and MOBIO [14] databases. BANCA database includes 82 subjects with equal number of males and females. We utilized the Grand Test (G) protocol which includes all available scenarios (controlled, degraded and adverse) and amounts to 3420 images for training and 3120 images for evaluation. On the other hand, in the MOBIO database, there are 150 subjects of which 99 is male and 51 is female. In order to decrease the database size, every fourth image is utilized in the experiments, resulting in 3870 and 1667 images for training and evaluation, respectively, with similar gender distribution to that of the whole database.

[1] http://www.idiap.ch/software/bob/

[2] Code available at: http://pypi.python.org/pypi/estimate.gender

TABLE I. NUMBER OF MALE AND FEMALE SUBJECTS AND IMAGES IN EACH DATABASE FOR TRAINING AND EVALUATION SETS

	Training		Evaluation	
Database	Male Sub./Img.	Female Sub./Img.	Male Sub./Img.	Female Sub./Img.
BANCA	28/1710	28/1710	13/1560	13/1560
MOBIO	61/2660	31/1210	38/1093	20/575
FERET	152/152	152/152	60/60	47/47
LFW (avg)	3410/8205	1189/2382	853/2051	297/595
MORPH-II (avg)	9199/37317	1733/6790	2323/9329	437/1698

Regarding to the E2 experiments, most of the existing work on gender classification do not use a standardized evaluation procedure which makes it very difficult to compare different methodologies. However recently, an evaluation protocol was proposed by Gehrig et al. [8] as one of the BeFIT (Benchmarking Facial Image Analysis Technologies) benchmarks which deals with two different scenarios: under controlled laboratory conditions and under uncontrolled real life conditions.

For benchmarking gender classification algorithms under controlled conditions, MORPH-II [19] database is proposed. This database consists of 55134 face images of 12012 male and 1605 female subjects. On the other hand, for uncontrolled scenario, the Labeled Faces in the Wild (LFW) [11] database is recommended which is composed of 13233 face images from 4263 male and 1486 female subjects.

For both of these databases, 5-fold cross-validation procedures are prepared and made available on BeFIT website[3]. In order to avoid learning the identity of the subjects in the training set instead of their genders, all images of a subject are ensured to appear in only one fold at each iteration. Moreover, the distribution of age, gender and ethnicity in the folds are kept similar to the distribution in the whole database.

Although FERET is one of the most commonly used databases for gender classification studies, it is not included in the BeFIT protocols due to its small size. Nevertheless, we include it in our experiments for the sake of comparability to previous works. To this end, a small partition of the FERET database (411 images) is utilized according to the evaluation protocol that was created and used by Mäkinen et al. in [13][4]. Unlike MORPH-II and LFW, the FERET protocol includes almost equal number of female and male images for training and evaluation.

The details of gender distribution in training and evaluation sets of all 5 databases used in our experiments are given in Table I.

In each of the 5 folds in MORPH-II and LFW databases and in the single fold of the FERET database, the classifiers are trained using the features extracted from training images and the success rates are measured and reported on the evaluation images for within-database experiments.

Finally, for the last set of experiments, E3, the training and evaluation steps are done on different databases in order to measure the generalizability of the classifiers. The same folds are used for comparability. For instance, the evaluation set of the FERET database is used to separately test the 5 classifiers trained by 5 training folds of LFW database and the result is averaged. The same is done for 5 evaluation sets from 5 folds of MORPH-II, resulting in 25 performance metrics, which are also averaged.

Additionally, in order to observe the effect of gender distributions in the training sets on the classification performances, E2 and E3 experiments for LFW and MORPH-II are repeated after the excess male samples are removed from each training step. In other words, the training sets of all folds are made gender-balanced for these experiments which are symbolized as LFW-B and MORPH-II-B.

B. Evaluation Metrics

In the BeFIT benchmark, three different evaluation metrics are given: *accuracy* (ACC), *average correct rate* (ACR) and *area under receiver-operator characteristic (ROC) curve* (AUC).

ACC is defined by:

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

where TP is the number of correctly classified positive (male) samples, TN is the number of correctly classified as negative (female) samples and P and N are total numbers of positive and negative samples. The main disadvantage of this metric is that for imbalanced databases and/or classifiers with different accuracies for male and female images, it can be highly misleading. To have a clearer understanding, *true positive* and *true negative rates* (TPR and TNR) can also be calculated:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{P}}, \quad \mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{N}} \tag{2}$$

ACR is the average of TPR and TNR which is more reliable than ACC in cases of database imbalance.

Finally, AUC measures the area under the ROC curve and reaches its maximum value of 1 if the system works perfectly. The main drawback of this metric is that it does not rely on a decision threshold, which hinders the generalizability measurements. For this reason, in our study, we will report the experimental results only using TPR, TNR, ACC and ACR.

C. Experimental results

The results of the experiments to find the optimal algorithm parameters and then to evaluate the final system within- and cross-database are given in the following subsections:

1) Optimal parameters: Once the E1 experiments are run and ACR values are calculated for all possible combinations of various cropping and LBPHS extraction parameters, the best configuration is found as the one in the top rank for both BANCA and MOBIO databases.

The results show that the highest rank for both databases is obtained when faces are cropped to size **105**×**70** with IOD of **30**, **modified** LBP codes are computed using **8** sampling points on a circle of radius **2** and the final features are constructed by concatenating histograms extracted from **12**×**12** non-overlapping blocks. With these parameters, 90.77% accuracy is obtained for BANCA, whereas it is 93.35% for MOBIO.

[3] http://fipa.cs.kit.edu/431.php

[4] http://www.sis.uta.fi/ em55910/datasets/

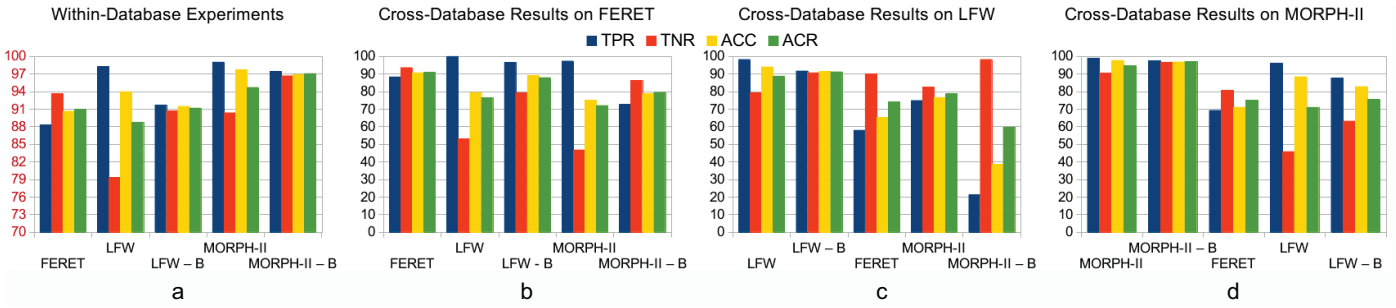

Fig. 3. TPR, TNR, ACC and ACR values for within-database experiments; a) for all databases and cross-database experiments for b) FERET, c) LFW and d) MORPHO-II with and without gender-balanced (B) training sets.

TABLE II. TPR, TNR, ACC AND ACR VALUES FOR E2 EXPERIMENTS. ADDITIONALLY, SOME PREVIOUSLY REPORTED RESULTS ARE GIVEN FOR COMPARISON OF WHICH THE ONES THAT UTILIZE THE SAME PROTOCOL ARE MARKED WITH AN ASTERISK (*)

	TPR	TNR	ACC	ACR
FERET	88.33%	93.62%	90.65%	90.98%
LFW	98.24%	79.31%	93.98%	88.78%
LFW-B	91.71%	90.66%	91.48%	91.19%
MORPH-II	99.03%	90.37%	97.69%	94.70%
MORPH-II-B	97.46%	96.68%	96.84%	97.07%
FERET* (LBP+SVM [1] - our imp.)	93.33%	85.11%	89.72%	88.22%
FERET* (LBP+SVM [1])	-	-	93.46%	-
FERET* (LBP+SVM [13])	-	-	82.06%	-
FERET* (Raw+SVM [13])	-	-	86.54%	-
LFW (MLBP+Adaboost+SVM [22])	95.98%	91.98%	94.4%	93.98%
LFW* (LBP+SVM [6])	97.01%	82.97%	93.83%	89.99%
LFW* (Gabor+SVM [6])	97.47%	82.16%	94.01%	89.82%
MORPH-II (BIF+KCCA [10])	-	-	98.45%	-

2) Within-database experiments: The results for the E2 experiments that are based on the findings of E1 are given in Figure 3-a.

For FERET, an ACR of 90.98% is achieved with the proposed method. The average ACR of 5 folds is found to be 88.78% for LFW and 94.70% for MORPH-II. On the other hand, after the gender distributions are balanced in the training sets, ACR for LFW increases to 91.19% and for MORPH-II to 97.07%. More detailed results are given in Table II.

3) Cross-database experiments: The results for the E3 experiments that are based on the findings of E1 are given in Figure 3-b,c,d. According to these results, the performances tend to drop when another database is used to train the classifiers. The result on the LFW is observed to be the most effected one when training sets of MORPH-II are used after gender distributions are adjusted (ACR of 59.73%).

The results reveal that most successful classifier in cross-database evaluations is the one trained with gender-balanced training sets of LFW. In fact, among all E3 experiments, the least effected result is obtained on FERET with this classifier, for which the ACR value is computed as 87.91%.

In general, utilization of the remarkably small FERET database for training, leads to a substantial deterioration on LFW and MORPH-II performances, with 74.02% and 74.98% ACR, respectively. In Table III, the results for the cross-database experiments are presented in further detail.

TABLE III. TPR, TNR, ACC AND ACR VALUES FOR E3 EXPERIMENTS. EVALUATION DATABASES ARE IN THE PARENTHESES.

	TPR	TNR	ACC	ACR
FERET (LFW)	57.99%	90.06%	65.20%	74.02%
FERET (MORPH-II)	69.30%	80.66%	71.05%	74.98%
LFW (FERET)	100.0%	53.19%	79.44%	76.60%
LFW (MORPH-II)	96.18%	45.82%	88.43%	71.00%
LFW-B (FERET)	96.67%	79.15%	88.97%	87.91%
LFW-B (MORPH-II)	87.71%	63.21%	82.81%	75.46%
MORPH-II (FERET)	97.33%	46.81%	75.14%	72.07%
MORPH-II (LFW)	74.90%	82.61%	76.64%	78.76%
MORPH-II-B (FERET)	72.67%	86.38%	78.69%	79.52%
MORPH-II-B (LFW)	21.22%	98.25%	38.55%	59.73%

D. Discussions

Due to the shape of the face, the aspect ratio (width:height) of the cropped image is most commonly taken as less than or equal to 1 in the literature. When we look at the selected cropping parameters that give the best results, we see that they are very different from the ones proposed in the previous studies because the width is larger than the height.

The final system that is based on the findings of the grid search for optimal algorithm parameters leads to comparable results to the state-of-the-art. For FERET, the best performances are reported by [1], however, our implementation of the same method leads to 89.72% ACC and 88.22% ACR which is marginally lower than the obtained results. For LFW, the ACC obtained in [22] is found to be the highest. On the other hand, the protocol used in that study is not the same as ours. In fact, most of the face images that are difficult to establish the ground truth or not near-frontal are omitted which simplifies the gender classification challenge and boosts the performance. The second best result on the same database is reported as 94.01% ACC in [6]. This is higher than the ACC achieved by the proposed method but when ACR values are compared, it is revealed that this is just a false appearance due to the strong gender imbalance in the evaluation set. In fact, the proposed method reaches a better result in terms of ACR. Finally, to the best of our knowledge, there are no gender classification studies yet which utilize MOPRH-II with the BeFIT protocol. In the only study that we could find with MORPH-II [10], the experiments are conducted in 3 folds and hence their results are not fully comparable to ours.

The gender-balanced tests on LFW and MORPH-II reveal that the gender distribution in the training sets has a huge impact on the performances. Since the number of male samples are much higher than females in both databases, the classifiers

are trained to have a tendency to label a test image as male. This is noticed clearly in the TNR values in Table II. When the training sets are balanced, TNRs increase remarkably, while TPRs decrease slightly. These experiments also help to realize the difference in ACC and ACR metrics. The true impact of the distribution uniformization can be observed in the ACR values, whereas it reflects negatively on the ACC due to the similar gender imbalance towards males in the evaluation sets.

Lastly, with the cross-database experiments, it has been shown that generalizability of the trained systems is a critical issue. The performances have worsen significantly compared to the within-database tests and TPR and TNR have become extremely imbalanced. ACR of the classifier trained with FERET roughly drops to 74% with LFW and MORPH-II evaluation sets. Similar decline in performances is also observed for classifiers trained with LFW and MORPH-II. Systems trained with gender-balanced training sets of LFW have the best generalization, since they bear the minimum loss.

V. Conclusion

Automated gender classification from face images is still an active topic today, especially in the case of uncontrolled real-world conditions. In this paper, we aim to contribute to the current state of research in this domain in two ways; firstly by analyzing the influence of cropping and LBPHS extraction parameters on the performances and selecting the optimal configuration and secondly by evaluating the final system using the public BeFIT protocol; using balanced and imbalanced training sets in within- and cross- database experiments.

The results show that the proposed system achieves a higher ACR than the best performance reported so far on the uncontrolled LFW database. It also gives comparable results on controlled FERET and MOPRH-II databases.

During our research, we encountered two main problems in comparing our results with the previous studies: misleading characteristic of the most commonly used ACC metric for imbalanced evaluation sets and unreproducible experimental results. For instance, the results reported in [22] could not be reproduced or comparable experiments cannot be conducted since the utilized protocol is not available. To alleviate these issues, both ACC and ACR metrics are utilized in our paper and the code to generate the reported results is made available.

For future work, our aim is to construct gender classification systems with better generalization properties. Cross-database experiments show that this is an issue still waiting to be handled for real-world applications.

Acknowledgment

This study was supported by the Swiss National Science Foundation under National Centre of Competence in Research (NCCR) on IM2 (*http://www.im2.ch*).

References

[1] L. A. Alexandre. Gender recognition: A multiscale decision fusion approach. *Pattern Recognition Letters*, 31(11):1422–1427, 2010.

[2] A. Anjos, L. E. Shafey, R. Wallace, M. Günther, C. McCool, and S. Marcel. Bob: a free signal processing and machine learning toolbox for researchers. In *ACM Conf. on Multimedia Systems, Japan*, 2012.

[3] E. Bailly-Bailliére, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Mariéthoz, J. Matas, K. Messer, V. Popovici, F. Porée, et al. The banca database and evaluation protocol. In *Audio-and Video-Based Biometric Person Authentication*, pages 625–638. Springer, 2003.

[4] W.-S. Chu, C.-R. Huang, and C.-S. Chen. Gender classification from unaligned facial images using support subspaces. *Information Sciences*, 221(0):98 – 109, 2013.

[5] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[6] P. Dago-Casas, D. González-Jiménez, L. L. Yu, and J. L. Alba-Castro. Single-and cross-database benchmarks for gender classification under unconstrained settings. In *IEEE Int. Conf. on Computer Vision Workshops*, pages 2152–2159, 2011.

[7] J. G. Daugman et al. Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *Optical Society of America, Journal, A: Optics and Image Science*, 2(7):1160–1169, 1985.

[8] T. Gehrig, M. Steiner, and H. Ekenel. Draft: Evaluation guidelines for gender classification and age estimation, 2011.

[9] B. A. Golomb, D. T. Lawrence, and T. J. Sejnowski. Sexnet: A neural network identifies sex from human faces. In *Conf. on Advances in Neural Information Processing Systems 3*, NIPS-3, pages 572–577. Morgan Kaufmann Publishers Inc., 1990.

[10] G. Guo and G. Mu. Joint estimation of age, gender and ethnicity: CCA vs. PLS. In *IEEE Int. Conf. and Workshops on Automatic Face and Gesture Recognition*, pages 1–6, 2013.

[11] G. B. Huang, M. Mattar, T. Berg, E. Learned-Miller, et al. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on Faces in'Real-Life'Images: Detection, Alignment, and Recognition*, 2008.

[12] A. Jain, J. Huang, and S. Fang. Gender identification using frontal facial images. In *IEEE Int. Conf. on Multimedia and Expo*, pages 4–pp, 2005.

[13] E. Makinen and R. Raisamo. Evaluation of gender classification methods with automatically detected and aligned faces. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 30(3):541–547, 2008.

[14] S. Marcel, C. M. Cool, C. Atanasoaei, F. Tarsetti, J. Pesán, P. Matejka, J. Cernocky, M. Helistekangas, and M. Turtinen. MOBIO: Mobile biometric face and speaker authentication. In *IEEE Conf. on Computer Vision and Pattern Recognition*, San Francisco, CA, USA, 2010.

[15] B. Moghaddam and M.-H. Yang. Learning gender with support faces. *Trans. Pattern Analysis and Machine Intelligence*, 24(5):707–711, 2002.

[16] T. Ojala, M. Pietikainen, and D. Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern recognition*, 29(1):51–59, 1996.

[17] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *Trans. Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002.

[18] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss. The feret evaluation methodology for face-recognition algorithms. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(10):1090–1104, 2000.

[19] K. Ricanek and T. Tesafaye. Morph: A longitudinal image database of normal adult age-progression. In *Int. Conf. on Automatic Face and Gesture Recognition*, pages 341–345, 2006.

[20] M. Riesenhuber and T. Poggio. Hierarchical models of object recognition in cortex. *Nature neuroscience*, 2(11):1019–1025, 1999.

[21] Á. Serrano, I. M. de Diego, C. Conde, and E. Cabello. Recent advances in face biometrics with gabor wavelets: A review. *Pattern Recognition Letters*, 31(5):372–381, 2010.

[22] C. Shan. Learning local binary patterns for gender classification on real-world face images. *Pattern Recognition Letters*, 33(4):431–437, 2012.

[23] N. Sun, W. Zheng, C. Sun, C. Zou, and L. Zhao. Gender classification based on boosting local binary pattern. In *Advances in Neural Networks*, pages 194–201. Springer, 2006.

[24] Z. Sun, G. Bebis, X. Yuan, and S. J. Louis. Genetic feature subset selection for gender classification: A comparison study. In *IEEE Workshop on Applications of Computer Vision*, pages 165–170, 2002.

[25] J. Trefnỳ and J. Matas. Extended set of local binary patterns for rapid object detection. In *Computer Vision Winter Workshop*, 2010.

[26] L. Wiskott, J.-M. Fellous, N. Krüger, and C. von der Malsburg. Face recognition and gender determination. In *Int. Workshop on Automatic Face and Gesture Recognition*, pages 92–97, 1995.

[27] T. Zhang and B.-L. Lu. Selecting optimal orientations of gabor wavelet filters for facial image analysis. In *Image and Signal Processing*, pages 218–227. Springer, 2010.

[28] J. Zheng and B.-L. Lu. A support vector machine classifier with automatic confidence and its application to gender classification. *Neurocomputing*, 74(11):1926 – 1935, 2011.