Face Gender Classification: A statistical study when neutral and distorted faces are combined for training and testing purposes

Yasmina Andreu, Pedro García-Sevilla, Ramón A. Mollineda

Dept. de Lenguajes y Sistemas Informáticos, Universitat Jaume I
12071 Castellón de la Plana, Spain

Abstract

This paper presents a thorough study of gender classification methodologies performing on neutral, expressive and partially occluded faces, when they are used in all possible arrangements of training and testing roles. A comprehensive comparison of two representation approaches (global and local), three types of features (grey levels, PCA and LBP), three classifiers (1-NN, PCA+LDA and SVM) and two performance measures (CCR and d') is provided over single- and cross-database experiments. Experiments revealed some interesting findings, which were supported by three non-parametric statistical tests: when training and test sets contain different types of faces, local models using the 1-NN rule outperform global approaches, even those using SVM classifiers; however, with the same type of faces, even if the acquisition conditions are diverse, the statistical tests could not reject the null hypothesis of equal performance of global SVMs and local 1-NNs.

Keywords: Face Analysis; Gender Classification; Global/Local Representation; Cross-database Experiment

1. Introduction

Classifying demographic traits such as gender, age and race is useful in countless tasks. In particular, gender classification can be applied to dynamic market studies, human-computer interaction, personalised services in a large number of businesses, among others.

In the area of face analysis, face recognition [1, 2, 3, 4, 5, 6] and facial expression analysis [7, 8, 9, 10] are extensively studied compared to the gender classification problem which is addressed less often [11, 12]. This could partly be due to a general belief that gender classification is similar to a face recognition problem with only two classes. To the best of our knowledge, there are

Email addresses: yandreu@uji.es (Yasmina Andreu), pgarcia@uji.es (Pedro García-Sevilla), mollined@uji.es (Ramón A. Mollineda)

not published studies that support this statement by exploring the performance differences of automatic systems when dealing with face recognition and gender classification. However, these studies are easily found in the psychology literature [13, 14, 15]. In [13], it is clearly stated that, in order to identify a face, the information that makes it unique has to be encoded. In contrast, to recognize the gender of a face the information encoded must be shared by a group of different faces (male or female). From the point of view of data complexity, gender classification is a 2-class problem with a generally large number of face images per class from different people resulting in sparse classes, while face recognition is a multi-class problem with usually very few faces per class that belong to the same individual. Therefore, gender classification problems have commonly a much higher intra-class variance than face recognition problems.

Many believe that systems designed to address a face recognition problem generally succeed in gender classification. After the explanation above, it seems clear that this is not always true. One example could be the well-known face recognition system proposed by Martinez [1]. It uses only one training sample per class. When addressing a gender classification problem, it is desirable to have a broad and diverse training data set so each gender could be as well characterised as possible. For this reason, this would not be the best approach to deal with gender classification.

Although the concept of gender is universally known, what type of information allows automatic systems to discriminate between male and female faces is not clear yet. According to psychological studies [16], humans use configural and featural information for recognizing faces, although it is possible for us to perform quite well in the absence of one of them. Since humans use both types of information when analysing faces, some authors have decided to use global (configural information) as well as local (featural information) descriptors assuming that it will ease the problem of classifying gender on automatic systems. Based on this idea, studies combining local and global features [17, 18] conclude that using both types of features provides better face characterisations and hence better classification rates than using just one of them. It should be noted that these studies used occlusion-free face images.

Most of the areas where automatic gender classification has interesting applications are usually set in real environments where the accessories and clothes worn by subjects are beyond our control, and people express their feelings through facial expressions. These are the main reasons why automatic gender classification systems should be able to properly classify expressive and partially occluded face images. Many studies have been published proposing several methodologies for recognizing faces in the presence of occlusions [1, 5, 6], as opposed to the very few published studies on gender classification of occluded faces [19]. Toews and Arbel [19] propose a methodology for classifying visual traits using the Object Class Invariant (OCI) model. Faces are described by an OCI consisting of a segment line from the bottom of the nose to the forehead and a set of model features denoted by scale-invariant geometric and appearance image information. Using images from the FERET database, the best classification rate is 83.7% obtained using a Bayesian classifier. In addition, the

authors test their OCI model for classifying gender from simulated occluded faces. That is, images from the FERET database with a resolution of 256×384 pixels were artificially obscured by a black circle of different radii. With an occlusion of radius 40 pixels, the classification rate is 75%, however when the occluding radius goes up to 80 pixels, the classification rate drops to 60% which is the rate of male faces in the data set.

In the current literature, most of the automatic gender classification systems use the same face database for obtaining the training and test sets [20, 21, 22]. In this case, the acquisition conditions of training and test images are exactly the same which is far from a realistic scenario. Bekios-Calfa et al. [23] proved that single-database experiments are optimistically biased and present a cross-database study on gender classification. Regarding cross-database experiments with a reasonable amount of training samples, SVM+RBF roughly achieves 80% of success. However, if there is less training data and a broad demography, all the compared classifiers achieve similar classification rates of around 70-75%. All three face databases used in that work contained non-occluded faces.

This paper presents an experimental study of gender classification from face images comparing two different representation approaches, three types of features and several classifiers using two performance measures. Experiments are carried out on single databases and crossing databases to explore more realistic scenarios. Furthermore, experiments focus on classifying the gender from face images showing different facial expressions and from partially occluded faces.

The main contributions of this paper are the following:

- A thorough experimental study of gender classification methodologies from neutral and distorted faces used in all possible combinations of training and testing roles. Distorted faces refer to faces affected by facial expressions or by real occlusions caused by wearing a scarf or sunglasses. Conclusions are supported by three statistical tests applied to two different performance measures.

- Solid conclusions are drawn on the strengths and weaknesses of two representations approaches (global and local), three types of features (grey levels, PCA and LBP) and three classifiers (1-NN, SVM and PCA+LDA) when addressing the problem of gender recognition.

- A reliable assessment of the robustness of the presented methods to changes in acquisition conditions and demographic variables, such as age and ethnicity, by performing cross-database experiments.

The rest of the paper is structured as follows: Section 2 outlines the methodology adopted for performing the experiments and details the processes of characterizing face images and the classifiers to be used; Section 3 presents the databases involved in the experiments; Section 4 describes the experimental set-up and the statistical tests which are applied for comparing the results; in Section 5 the results of the experiments are discussed and the statistical differences found among the performances of the different classification models are provided; finally, the conclusions are presented in Section 6.

2. Methodology

An overview of the methodology adopted for carrying out the experiments is given before detailing each step of this process:

Preprocessing of face images is necessary to isolate the area of the image containing the face, to normalise the contrast of the image, and to suppress illumination problems. First, the face is detected automatically using the Viola and Jones algorithm [24] implemented in the OpenCV [25] library. Next, the area of the image containing the face is extracted and then equalized and resized. The interpolation process required for resizing the image uses a three-lobed Lanczos windowed sinc function [26] which keeps the original image aspect ratio. In the end, an equalized face image is passed to the feature extraction step. It should be noted that no technique for aligning faces is applied, so unaligned faces will be classified.

Feature extraction is applied to each preprocessed face image to characterise the face using feature vectors. See further details in Section 2.2.

Classification follows a standard scheme: a model trained from previously seen faces predicts the gender of a test face. This process is explained in Section 2.3.

Performance assessment is carried out using two different measures (correct classification rates and d-prime). These two measures are detailed in Section 2.4.

2.1. Representation Approaches

Two representation approaches are compared in this experimental study: a global and a local scheme.

In the global approach, the face is characterised as a whole. Therefore, this representation provides configural as well as featural information.

In the local approach, the face is characterised by a collection of local regions which provide information about the appearance of each part of the face in isolation. A series of overlapping patches of $M \times M$ pixels are considered over the area of the image where the face was detected. From one patch to its neighbour, there is a one pixel shift. Given a position (i, j) in a face image, we define $D_{i,j}$ as the neighbourhood of patches associated to (i, j). For a given patch $p_{k,l}$, centred at position (k, l), $p_{k,l} \in D_{i,j}$ iff $|i - k| \leq P$ and $|j - l| \leq P$. The constant P defines the size of the neighbourhood. In the classification step, the class of each patch is predicted by the classifier constructed using its neighbouring patches in the training set. This local classification scheme based on neighbourhoods was designed to have a higher tolerance towards distortions in face images, inaccurate face detections and alignments.

2.2. Features

This study compares three types of features: grey levels (raw information), Principal Component Analysis of the grey level space (transformed information) and Local Binary Patterns (texture information), all combined with the two aforementioned representation approaches.

Regardless of which feature is going to describe the face, the method for extracting features depends on which representation approach (global or local) is adopted. For global features, the method considers all pixels within the area of the image where the face was detected, resulting in one feature vector that describes the face. For local features, the method is applied to each one of the patches extracted from the image. Consequently, several feature vectors describe one face. The feature extraction methods are explained below.

2.2.1. Grey Levels

The grey level values of the pixels forming the area of interest in the image (the whole face or a local patch) are vectorised resulting in a feature vector.

For global features, the face is characterised by a feature vector of the grey level values of the pixels within the area where the face was detected. For local features, the face is characterised by a set of feature vectors of a common length equal to the number of pixels in each patch. In this case, the face is characterised by as many feature vectors as patches are in the face image.

2.2.2. Principal Component Analysis

Principal Component Analysis (PCA) [27] searches for a subspace whose basis vectors correspond to those directions in the original space with maximum variance. Let W be a linear transformation that maps the original $d-$dimensional space onto a $f-$dimensional feature subspace. Then, new feature vectors $y_i \in \Re^f$ are defined by $y_i = W^T x_i$ where $x_i \in \Re^d$. This method has been widely used in face recognition [4, 5, 28] leading to good recognition accuracies. Those results have encouraged some researchers to classify gender in the transformed PCA space [29] where very good performances are reported.

In this work, PCA features are extracted from the whole face area (global approach) and from the neighbourhood of each position (local approach). In both cases, the target subspace is built from training face images by retaining those eigenvectors accounting for 95% of the variance. For obtaining global features, the PCA basis are calculated from the grey level values of the area of the image where the face was detected. Then, the PCA transformation is applied to each face image resulting in a feature vector which describes it. For obtaining local features, the PCA basis are calculated separately for each patch neighbourhood.

2.2.3. Local Binary Patterns

Local Binary Patterns (LBP) were originally defined to characterise image textures [30]; more recently, they have been used as face descriptors [3]. A binary number describes each pixel in the image and it is calculated considering

a neighbourhood around each pixel. Then, all neighbours are given either value 1, if they are brighter than the central pixel, or value 0 otherwise. The values assigned to the neighbours are read sequentially in the clockwise direction to form the binary pattern which characterises the central pixel. In the end, the image is characterised by a histogram of the LBP values of all the pixels. This description was improved by using the so-called uniform LBPs [31]. The uniform patterns have at most two one-to-zero or zero-to-one transitions in the circular binary code.

In this work, uniform LBPs with neighbourhoods of radius 2 and 8 sample points are used. As a result, a histogram of 59 bins is extracted where 58 bins correspond to all the possible uniform LBPs and the extra bin is for accumulating the non-uniform patterns.

For global features, instead of characterising the whole face, several non-overlapping regions of the face image are described separately; this is how LBP features have been successfully used to represent facial images [3]. For each region, a histogram of LBP features is extracted and then, all histograms are concatenated to form one feature vector as a global description of the face. For local features, each patch is characterised by its corresponding histogram of LBPs.

2.3. Classifiers

Three different types of classifiers are used in this study: Nearest Neighbour, PCA+LDA and Support Vector Machines.

Following a global approach, a standard classification scheme is adopted where a classifier previously trained with the features extracted from the training face images predicts the gender of a test face. In the local approach, the classification process is slightly different and it works as follows. Let $C_{i,j}$ be the local classifier trained with $D_{i,j}^{tra}$, the set of training patches within the neighbourhood of the position (i, j) in the face images. Given the set of patches extracted from a test image, a class label for patch $p_{i,j}$ is predicted by $C_{i,j}$ $\forall i, j$, resulting in N predicted classes. Finally, the predicted gender of the face is obtained by a majority vote of the N local predictions.

2.3.1. Nearest Neighbour (1-NN)

The intuition underlying nearest neighbour classification is quite straightforward. A test sample is classified by its nearest neighbour in the training set. The metric used is the Euclidean distance.

Global 1-NN works as the well-known 1-NN classifier.

Local 1-NN consists of a set of local classifiers. In total, there are as many 1-NN classifiers as patches. Each one of these local classifiers searches for the nearest neighbour only among the patches within the corresponding neighbourhood.

2.3.2. Principal Components + Linear Discriminant Analysis (PCA+LDA)

Linear Discriminant Analysis (LDA) [27] is a method to find a linear combination of features that best discriminate among classes. It looks for those features that enlarge the difference of the class means. In other words, it searches for a feature space where classes are better separated.

LDA is most commonly applied to a lower-dimensional intermediate space in order to avoid the mathematical problems due to the fact that the number of samples per class is usually small with respect to the dimensionality of the original feature space (for details see [32]). Hence, in gender classification problems, LDA is usually applied after reducing the dimensionality with PCA [23].

As a result, LDA provides a subspace of at most $c-1$ dimensions where c is the number of classes. In a gender classification problem there are two classes, so LDA provides a $1-$dimensional feature space. An optimal binary partition of the $1-$dimensional feature space is searched and used as a classification rule.

Global PCA+LDA works as the general PCA+LDA explained above.

Local PCA+LDA consists of a set of local classifiers, one PCA+LDA classifier per patch. Each one of these local classifiers uses the feature vectors extracted from the patches that belong to the corresponding neighbourhood.

2.3.3. Support Vector Machine (SVM)

A Support Vector Machine [33] constructs a hyperplane for an optimal class separation. There are many hyperplanes which can classify the data. Intuitively, the best separation can be achieved by the hyperplane that maximises margins to the nearest points of both classes. This classifier has been extensively used in many automatic facial analysis tasks, including gender classification [11, 12].

However, this classifier suffers from a high computational cost when provided with a large number of training samples. For this reason, we found that building as many local SVMs as patches are in the image was not computationally affordable. Therefore, in the experiments, SVM only adopts a global approach.

Global SVM works as a standard SVM. Particularly, the SVM implementation with a third degree polynomial kernel provided with LIBSVM 3.0 [34] is used in the experiments. After an exploratory study, this third degree polynomial kernel was chosen for its good ratio of computational cost to classification accuracy.

2.4. Classification Assessment

In order to assess the performance of classification models, two measures are used: Correct Classification Rate (CCR) and d-prime (d'). CCR, probably the most popular way of evaluating the effectiveness of a classifier, computes the percentage of correctly classified samples over the total number of samples. However, empirical and analytical evidence show that CCR may be strongly biased with regard to data imbalance, which could lead to inaccurate conclusions.

The second measure, d-prime [35], is a suitable approach for assessing the classifier behaviour on a two-class problem, because it is robust to skewed classes. Its computation is based on two performance indices computed separately on the two classes, which penalises biased classification results towards one of the classes. Thus, as compared to CCR, d' provides a different perspective to the analysis of classifier performance.

3. Description of the Face Databases

Three face databases are used in the experiments, two of them containing only neutral faces, and a third database containing neutral, expressive, and realistically occluded faces.

FERET (Facial Recognition Technology Database) [35] contains 12,922 colour images of 512×768 pixels corresponding to 994 people's faces ranging from ages 10 to 70 and from different races. Specifically, it contains face images of 412 subjects of less than 20 years old, 442 adults aged 20-40 and 140 adults aged 50-70. There are 225 Asian faces, 73 African-American faces, 57 Hispanic faces, 618 Caucasian faces and 21 from other races. There were 13 face images collected from each subject with different face poses turning the head right or left with several degrees.

Our experiments use only the images corresponding to frontal views of the face. The total number of images used from this database is 2,015 face images of 1,173 male and 841 female faces corresponding to 787 different subjects (427 males and 360 females).

PAL (Productive Aging Lab Face) [36] contains 575 colour images of size 640×480 pixels corresponding to 575 individuals (there is only one image per individual) with ages ranging from 18 to 93. It contains face images of 107 females and 115 males aged 18-29, 47 females and 29 males aged 30-49, 90 females and 33 males aged 50-69 and 106 females and 47 males aged 70-94. There are 89 African-American faces (26 males and 63 females), 434 Caucasian faces (158 males and 275 females) and 52 from other races (40 males and 12 females).

All the faces images from this database are used in the experiments.

AR [37] contains around 4,000 colour images of 768×576 pixels corresponding to 130 people's faces. Images feature frontal view faces with different facial expressions, illumination conditions, and occlusions (sunglasses and scarf). Information about the age and race of the subjects is not provided, although after majority sampling the database it can be said that all the individuals are young Caucasian adults.

Our experiments use the images corresponding to smile, angry, "screaming" and neutral facial expression, top occlusions caused by sunglasses and bottom occlusions caused by wearing a scarf. A total of 774 images

Global		Local	
(1)	1NN-grey-G	(8)	1NN-grey-L
(2)	1NN-pca-G	(9)	1NN-pca-L
(3)	1NN-lbp-G	(10)	1NN-lbp-L
(4)	PCALDA-G	(11)	PCALDA-L
(5)	SVM-grey-G		
(6)	SVM-pca-G		
(7)	SVM-lbp-G		

Abbreviation: `Classifier-Feature-G/L`

Table 1: Classification models considered in the experiments.

are used from this database. These images consist of 130 face images (74 males and 56 females) for each facial expression, 129 top occluded face images (75 males and 54 females) and 125 bottom occluded face images (72 males and 53 females).

The class balances are roughly 60% male and 40% female faces for FERET and AR, and 40%-60% for PAL. These class balances are more accurately provided for each data set in Table 2(a).

4. Experimental Set-up

The experimental set-up was designed to study the effectiveness of local and global representations, and grey levels, PCA and LBP face characterisations using 1-NN, PCA+LDA and SVM for classifying gender when distorted face images are used for training, testing or both. To the best of our knowledge, the problem of assessing the consequences of including non-neutral and occluded faces in the training and the evaluation of classifiers has not been extensively addressed in previous works. Additionally, to recreate realistic conditions, cross-database experiments are provided involving different databases for training and testing. A combination of a representation approach (global or local), a type of feature (grey levels, PCA or LBP) and a particular classifier (1-NN, PCA+LDA or SVM) will be referred to as a *classification model*. Table 1 enumerates the eleven different classification models considered in the experiments.

A summary of all the combinations of training-test data sets is shown in Table 2(a) (four letters A, B, C and D are used in this table to distinguish between experiments containing/not containing distortions in the training/test sets). In this table the class balances are also provided. Face images from the AR database are selected to form three data sets with an increasing level of difficulty: *AR neutral* contains only neutral faces, *AR light distortions* contains neutral and expressive faces, and *AR heavy distortions* contains the images of the previous data set and also occluded faces.

In order to provide statistical support for the existence of performance differences among classification models, several non-parametric statistical tests are

		Training data sets				
		FERET	PAL	AR neutral	AR light distortions	AR heavy distortions
		58:42	61:39	43:57	43:57	43:57
Test data sets	FERET	A	A	A	B	B
	PAL	A	A	A	B	B
	AR Neutral	A	A	A	B	B
	AR light distortions	C	C	C	D	D
	AR heavy distortions	C	C	C	D	D

(a) Combinations of training and test data sets (A: Training and test without distortions. B: Training with distortions and test without them. C: Training without distortions and test with them. D: Training and test with distortions). Class balances (percentage of male:female faces) are shown below each training data set.

		Training		
		With and Without Distortions	With Distortions	Without Distortions
Test data sets	With and Without Distortions	**G1**: A∪B∪C∪D	**G2**: B∪D	**G3**: A∪C
	With Distortions	**G4**: C∪D	**G5**: D	**G6**: C
	Without Distortions	**G7**: A∪B	**G8**: B	**G9**: A

(b) Definition of groups of experiments (from G1 to G9). These groups have been designed for a later statistical study of the classification rates of the experiments in each group.

Table 2: Summary of the data sets used in the different groups of experiments.

applied to two performance measures (Correct classification rate and d-prime). This statistical study of the results obtained will be performed over groups of experiments which are built regarding some distortion criteria to assess specific experimental scenarios. Table 2(b) shows the experiments that form each group.

When extracting grey levels and PCA features, for global descriptions, the area of the image containing the face is 45×36 pixels, giving a global feature vector length of 1620. For local representations, a face image is described by a total of 1170 local 49-dimensional feature vectors obtained from overlapping patches of size 7×7 pixels. When extracting LBP features, global characterisations consider 20 non-overlapping regions of 9×9 pixels from which LBP histograms are extracted and then concatenated to form a global feature vector of 1180 elements. For local LBP descriptions, a face is represented by 1170 feature vectors of 59 elements extracted from overlapping patches of 7×7 pixels. Local LBP features were also extracted from larger patches to test if the patch size influenced the classification task, concluding that the performance was not

strongly affected. Therefore, the size of patches is kept at 7×7 pixels to allow a direct comparison with the rest of the local features.

To assess classifier performances in single-database experiments, that is, experiments where the training and test sets are extracted from the same database, 5 repetitions of a 5-fold cross validation technique are executed (25 runs in total). The partitions needed for conducting these experiments were based on subjects instead of images. Therefore, images of the same individual could only be found in the training or the test set, but never in both. In cross-database experiments, only one simulation is performed, training with one database and testing with the other. For the local classification, each neighbourhood spans $P = 2$ positions in each direction from its center; hence, a neighbourhood covers 25 patches.

4.1. Statistical Tests

Due to the large number of experiments, a detailed comparison of performance is difficult. In order to ease this task, several tests[1] are applied to show whether statistical differences exist among the performances of the classification models. All of these statistical tests are based on a null hypothesis which states that all classification models perform equally. This is assumed certain and evidence is searched for in the data to reject it.

Firstly, Iman-Davenport's test [39] is computed to detect differences among the performances of all classifiers. This statistic is obtained from the equation:

$$F_F = \frac{(n-1)\chi_F^2}{n(k-1) - \chi_F^2} \tag{1}$$

which is compared to a F-distribution with $k-1$ and $(k-1)(n-1)$ degrees of freedom, where n denotes the total number of experiments, k the amount of classification models and χ_F^2 is the value of the Chi-square distribution with F degrees of freedom. In order to reject the null hypothesis, the F_F statistic should be higher than the corresponding value of the F-distribution. In that case, significant differences among the classification model performances exist.

Secondly, Holm's method [40] is applied to identify statistical differences between the most significant classification model and the remaining models. Holm's null hypothesis assumes that the performance of the former is statistically equal to the performance of the other models. Several pairwise hypotheses are checked sequentially, one per each of the models except for the most significant one. For a given significance level α, Holm's method checks if $P_{(i)} < \frac{\alpha}{k-i}$ where $P_{(i)}$ is the P-value of the i^{th} hypothesis and k the amount of classification models. If the condition is met, the i^{th} null hypothesis is rejected (i.e. the i^{th} model performs statistically worse than the most significant one).

Thirdly, Wilcoxon's Signed Rank test [41] provides pairwise comparisons, so statistical differences between each pair of classification models can be found.

[1] These statistical tests were conducted using KEEL data mining software [38].

For each pair, Wilcoxon's null hypothesis assumes that both classification models perform equally. This test proceeds by ranking the differences in performance of two models. Let d_i be the difference between the performances of two classification models on the i-th experiment. Then, the differences $d_i \ \forall i$ are ranked according to their absolute values. Let R_+ be the sum of the ranks where the 1st model outperforms the 2nd and R_- be the sum of the opposite cases. The ranks where $d_i = 0$ are split evenly among R_+ and R_-. When there is an odd number of cases where $d_i = 0$, one of those ranks is ignored. Being $Z = min(R_+, R_-)$, if Z is less or equal than the Wilcoxon distribution for n degrees of freedom, then the null hypothesis stating that both classification models perform equally well is rejected.

5. Discussion of the Experimental Results

In this section, we present a wide comparison of the performance of all the classification models involved in this study. In order to provide a comprehensive analysis, we carried out three statistical tests to the nine groups of experiments detailed in Table 2(b) considering both performance measures, CCR and d' (the performance results can be seen in Appendices A and B). Therefore, eighteen groups of experiments are discussed for each statistical test.

According to Iman-Davenport's statistic, significant differences exist among the performance of all classification models using both measures. To better grasps these differences, the results of Holm's and Wilcoxon's tests are discussed.

Holm's method

Holm's method results for CCR are shown in Figure 1 and for d' in Figure 2. The null hypotheses (which assume statistical equality) associated to those models above the double lines were rejected when compared with the most significant classification model (shown at the bottom of each table in Figures 1 and 2) with a 95% significance level.

Taking into account the statistical tests over both measures, global 1-NNs are always rejected using any type of features and PCA+LDA models, global and local, are rejected in most cases (except once using CCR and three times using d' out of the eighteen groups of experiments). In general, global SVMs and local 1NNs perform statistically better than the rest.

Regarding LBP features, the main difference between both measures is that, according to d', global SVMs and local 1-NNs are among the statistically superior models. However, this is not the case when using CCR. This suggests that LBP-based methods lead to more balanced performance rates between classes.

Wilcoxon's signed rank test

A summary of Wilcoxon's signed rank test for CCR is shown in Figure 3 and for d' in Figure 4. In these figures, the symbol "$\bullet$" indicates that the classification model in the row significantly outperforms the model in the column, and the symbol "$\circ$" indicates that the classification model in the column significantly

Training Dataset

Test Dataset	With & Without Distortions	With Distortions	Without Distortions
With & Without Distortions	**G1**	**G2**	**G3**
	1NN-lbp-G 0.005	1NN-lbp-G 0.005	1NN-lbp-G 0.005
	PCALDA-L 0.006	PCALDA-L 0.006	PCALDA-L 0.006
	1NN-pca-G 0.006	1NN-grey-G 0.006	1NN-pca-G 0.006
	1NN-grey-G 0.007	1NN-pca-G 0.007	1NN-lbp-L 0.007
	1NN-lbp-L 0.008	PCALDA-G 0.008	1NN-grey-G 0.008
	PCALDA-G 0.010	SVM-lbp-G 0.010	PCALDA-G 0.010
	SVM-lbp-G 0.013	1NN-lbp-L 0.012	1NN-pca-L 0.012
	1NN-pca-L 0.017	1NN-pca-L 0.017	SVM-pca-G 0.017
	SVM-pca-G 0.025	SVM-pca-G 0.025	SVM-lbp-G 0.025
	SVM-grey-G 0.050	SVM-grey-G 0.050	SVM-grey-G 0.050
	1NN-grey-L	**1NN-grey-L**	**1NN-grey-L**
With Distortions	**G4**	**G5**	**G6**
	1NN-lbp-G 0.005	PCALDA-L 0.005	1NN-lbp-G 0.005
	PCALDA-L 0.006	1NN-lbp-G 0.006	PCALDA-G 0.006
	1NN-pca-G 0.006	1NN-pca-G 0.006	PCALDA-L 0.006
	PCALDA-G 0.007	1NN-grey-G 0.007	1NN-lbp-L 0.007
	1NN-lbp-L 0.008	1NN-lbp-L 0.008	1NN-pca-G 0.008
	1NN-grey-G 0.010	PCALDA-G 0.010	1NN-grey-G 0.010
	SVM-lbp-G 0.012	SVM-lbp-G 0.012	SVM-grey-G 0.012
	SVM-grey-G 0.017	1NN-pca-L 0.017	SVM-pca-G 0.017
	1NN-pca-L 0.025	1NN-grey-L 0.025	SVM-lbp-G 0.025
	SVM-pca-G 0.050	SVM-grey-G 0.050	1NN-pca-L 0.050
	1NN-grey-L	**SVM-pca-G**	**1NN-grey-L**
Without Distortions	**G7**	**G8**	**G9**
	1NN-lbp-G 0.005	1NN-lbp-G 0.005	1NN-lbp-G 0.005
	PCALDA-L 0.006	PCALDA-L 0.006	1NN-pca-G 0.006
	1NN-grey-G 0.006	1NN-grey-G 0.006	1NN-grey-G 0.006
	1NN-pca-G 0.007	1NN-pca-G 0.007	PCALDA-L 0.007
	1NN-lbp-L 0.008	PCALDA-G 0.008	1NN-lbp-L 0.008
	PCALDA-G 0.010	SVM-lbp-G 0.010	PCALDA-G 0.010
	SVM-lbp-G 0.012	1NN-lbp-L 0.012	1NN-pca-L 0.012
	1NN-pca-L 0.017	SVM-pca-G 0.017	1NN-grey-L 0.017
	SVM-pca-G 0.025	1NN-pca-L 0.025	SVM-pca-G 0.025
	1NN-grey-L 0.050	SVM-grey-G 0.050	SVM-lbp-G 0.050
	SVM-grey-G	**1NN-grey-L**	**SVM-grey-G**

Figure 1: Holm's method results for measure CCR (correct classification rate) of all classification models compared with the most significant in each case (showed in bold at the bottom of each table) with a 95% significance level. All the models above the double line performed significantly worse than the model in bold. Refer to Table 2 for the description of the groups of experiments.

Training Dataset

Test Dataset	With & Without Distortions	With Distortions	Without Distortions
	G1	**G2**	**G3**
With & Without Distortions	1NN-lbp-G 0.005	1NN-lbp-G 0.005	1NN-lbp-G 0.005
	1NN-grey-G 0.006	PCALDA-L 0.006	1NN-pca-G 0.006
	1NN-pca-G 0.006	1NN-grey-G 0.006	1NN-grey-G 0.006
	PCALDA-L 0.007	1NN-pca-G 0.007	PCALDA-G 0.007
	PCALDA-G 0.008	PCALDA-G 0.008	PCALDA-L 0.008
	SVM-lbp-G 0.010	SVM-lbp-G 0.010	SVM-pca-G 0.010
	SVM-pca-G 0.012	SVM-pca-G 0.012	SVM-grey-G 0.012
	SVM-grey-G 0.017	SVM-grey-G 0.017	SVM-lbp-G 0.017
	1NN-lbp-L 0.025	1NN-lbp-L 0.025	1NN-lbp-L 0.025
	1NN-pca-L 0.050	1NN-pca-L 0.050	1NN-grey-L 0.050
	1NN-grey-L	**1NN-grey-L**	**1NN-pca-L**
	G4	**G5**	**G6**
With Distortions	1NN-lbp-G 0.005	PCALDA-L 0.005	1NN-lbp-G 0.005
	PCALDA-G 0.006	1NN-lbp-G 0.006	PCALDA-G 0.006
	1NN-pca-G 0.006	1NN-grey-G 0.006	1NN-pca-G 0.006
	1NN-grey-G 0.007	1NN-pca-G 0.007	1NN-grey-G 0.007
	PCALDA-L 0.008	PCALDA-G 0.008	SVM-pca-G 0.008
	SVM-pca-G 0.010	SVM-lbp-G 0.010	SVM-grey-G 0.010
	SVM-lbp-G 0.012	SVM-pca-G 0.012	PCALDA-L 0.012
	SVM-grey-G 0.017	1NN-pca-L 0.017	SVM-lbp-G 0.017
	1NN-lbp-L 0.025	1NN-lbp-L 0.025	1NN-lbp-L 0.025
	1NN-pca-L 0.050	1NN-grey-L 0.050	1NN-grey-L 0.050
	1NN-grey-L	**SVM-grey-G**	**1NN-pca-L**
	G7	**G8**	**G9**
Without Distortions	1NN-lbp-G 0.005	1NN-lbp-G 0.005	1NN-grey-G 0.005
	1NN-grey-G 0.006	PCALDA-L 0.006	1NN-pca-G 0.006
	PCALDA-L 0.006	1NN-grey-G 0.006	1NN-lbp-G 0.006
	1NN-pca-G 0.007	SVM-lbp-G 0.007	PCALDA-L 0.007
	PCALDA-G 0.008	1NN-pca-G 0.008	PCALDA-G 0.008
	SVM-lbp-G 0.010	PCALDA-G 0.010	1NN-lbp-L 0.010
	SVM-pca-G 0.012	SVM-pca-G 0.012	SVM-pca-G 0.012
	1NN-lbp-L 0.017	SVM-grey-G 0.017	SVM-lbp-G 0.017
	SVM-grey-G 0.025	1NN-lbp-L 0.025	SVM-grey-G 0.025
	1NN-grey-L 0.050	1NN-pca-L 0.050	1NN-grey-L 0.050
	1NN-pca-L	**1NN-grey-L**	**1NN-pca-L**

Figure 2: Holm's method results for measure d' of all classification models compared with the most significant in each case (showed in bold at the bottom of each table) with a 95% significance level. All the models above the double line performed significantly worse than the model in bold. Refer to Table 2 for the description of the groups of experiments.

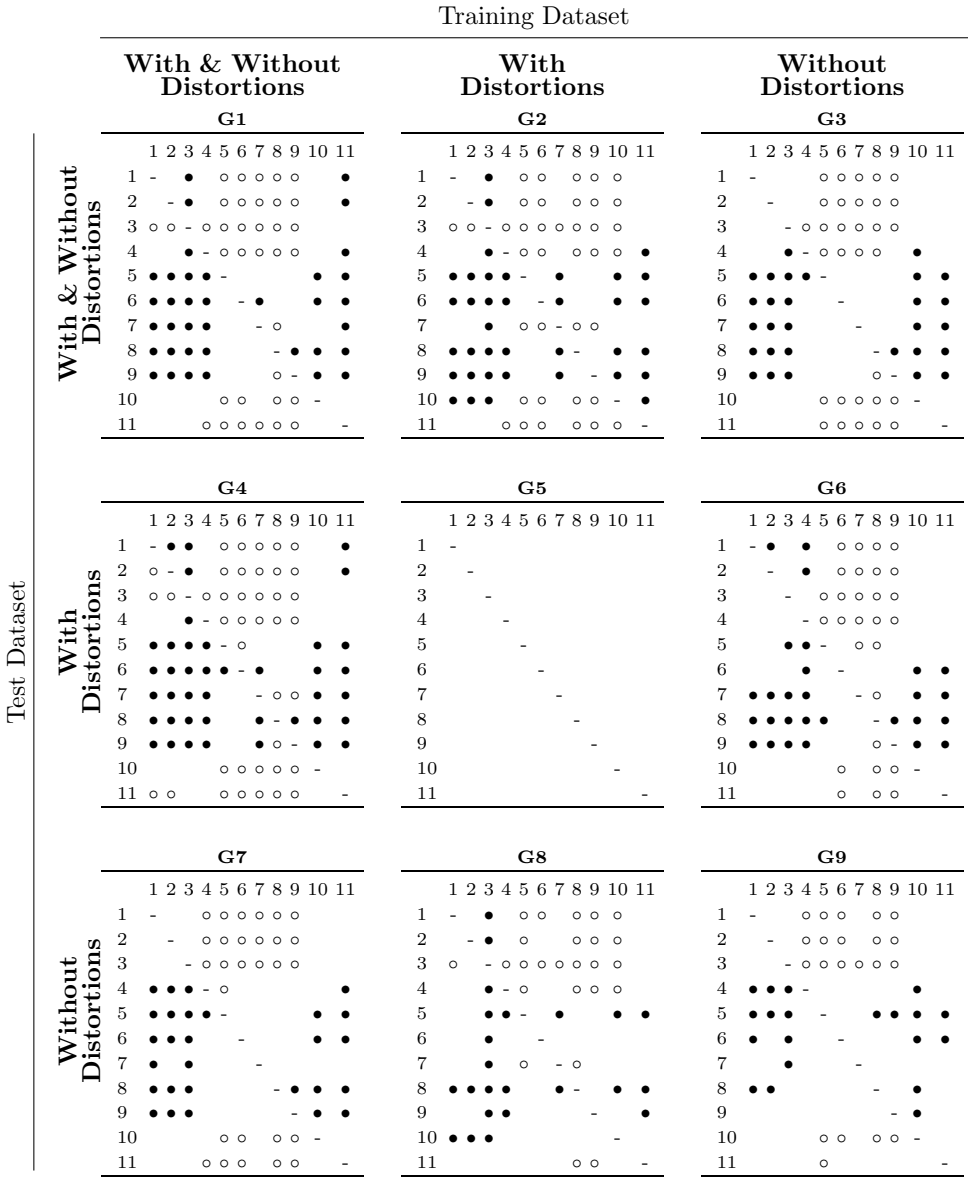

Figure 3: Summary of the Wilcoxon's statistic applied over CCR values for all pairs of classification models (above the main diagonal 90% confidence level, and below it 95%). The symbol "•" indicates that the classification model in the row significantly outperforms the model in the column, and the symbol "○" indicates that the model in the column significantly surpasses the model in the row. Refer to Table 1 for the description of the classification models numbered from 1 to 11 and to Table 2 for the description of the groups of experiments. In group G5, Wilcoxon's cannot find differences due to insufficient data.

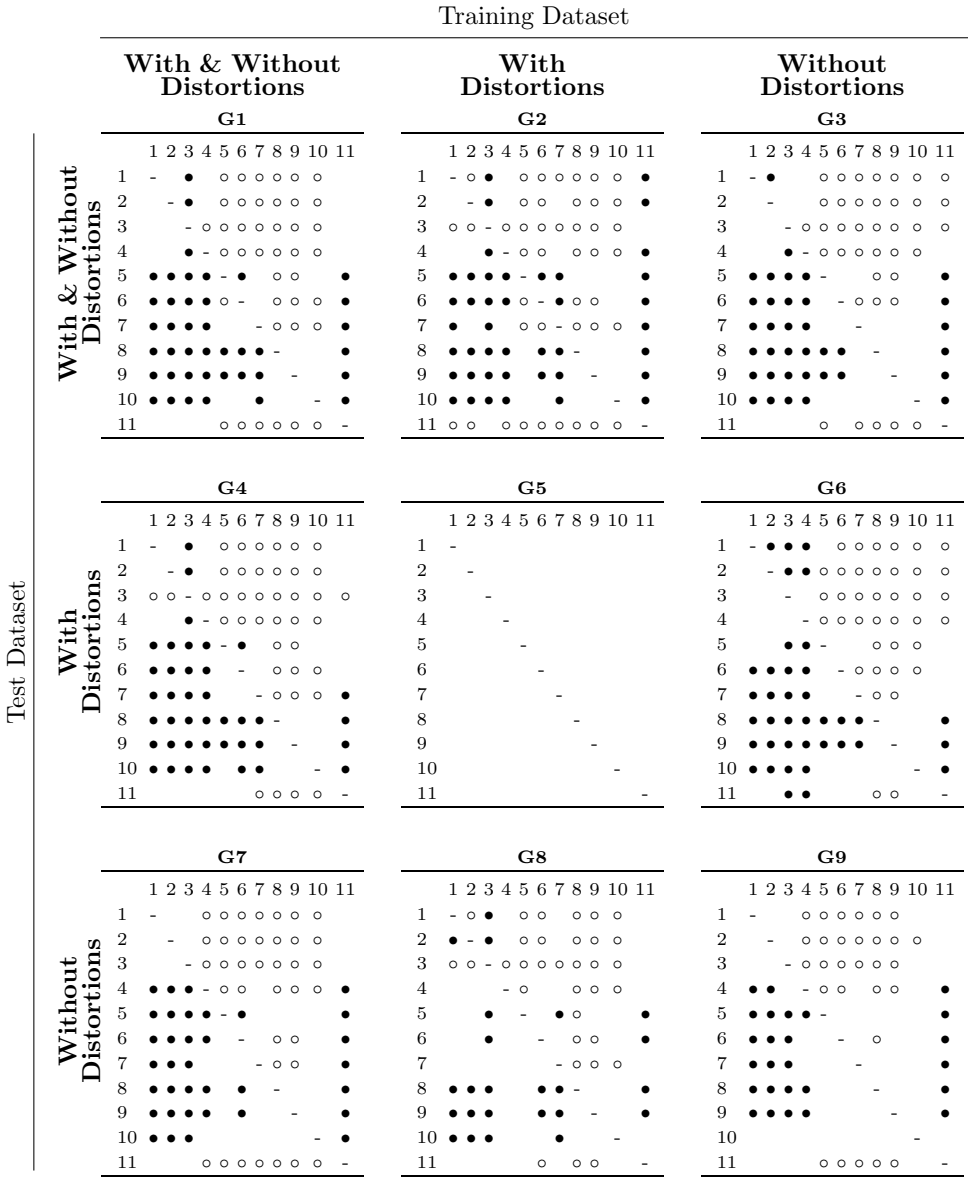

Figure 4: Summary of the Wilcoxon's statistic applied over d' values for all pairs of classification models (above the main diagonal 90% confidence level, and below it 95%). The symbol "•" indicates that the classification model in the row significantly outperforms the model in the column, and the symbol "○" indicates that the model in the column significantly surpasses the model in the row. Refer to Table 1 for the description of the classification models numbered from 1 to 11 and to Table 2 for the description of the groups of experiments. In group G5, Wilcoxon's cannot find differences due to insufficient data.

surpasses the model in the row. Above the main diagonal, the confidence level is 90% and, below it, it is 95%. The discussion of Wilcoxon's results is presented by groups of experiments (for details about the groups see Table 2(b)).

Groups G1, G2, G3, G4, and G7: Data sets of images with and without distortions were used for training or for testing. Wilcoxon's tests show that global SVMs and local 1-NNs statistically outperform the rest of the classification models, supporting the conclusion extracted from Holm's method. This is more clearly shown with the measure d' than with CCR, since with the latter the local 1-NN using LBP features does not always outperform all the other models. Besides, d'-based tests show that local 1-NN with grey levels and PCA features surpass the performance of global SVMs in most cases.

Group G5: Only distorted faces were used both for training and testing. Wilcoxon's test cannot find differences among the performances of the classification models due to insufficient data; to approximate a normal distribution, six or more experiments are needed [42]. Note that, in this group of experiments, Holm's method finds reasonable statistical differences since having a small number of experiments can only cause the non rejection of false null hypotheses [43].

Groups G6 and G8: Distorted faces were used for training and non-distorted for testing and vice versa. Wilcoxon's tests find that there is a tendency for local models to be statistically superior to global solutions. This conclusion is strongly supported by d'-based tests where local 1NNs perform statistically better than the rest of the models with a 90% confidence and than most of the models with a 95%.

Group G9: Only non-distorted faces were used both for training and testing. Wilcoxon's test findings differ depending on the measure used (d' or CCR). In d'-based tests, the six models (global SVMs and local 1NNs) that stand out in the other groups of experiments show a statistical superiority to the rest of global models (with the exception of local 1NNs using LBP features). In CCR-based tests, global SVM with grey levels seems to statistically outperform most of the other models. However, when these CCR statistical tests are applied only to the experiments of this group that use different databases for training and testing (that excludes the three single-database experiments), Iman-Davenport's statistic cannot find differences, Holm's method only rejects global 1-NN with PCA features and both, global and local, 1-NN with LBP features. In addition, Wilcoxon's test finds just a few statistical differences with a 95% of significance level, showing that global SVM with grey levels statistically outperforms only global 1-NNs with grey levels and both 1-NNs with LBP features (see Figure 5). These findings suggest that the statistical differences obtained for this group of experiments when using CCR measure should be mainly attributed to the experiments that use the same database for training and testing.

In general, looking at the results, there are two differentiated sets of classification models: global SVMs and local 1-NNs, and the rest. The former set undoubtedly obtains better performances than the latter in most groups of experiments. In cases where training faces present distortions and test faces do

Holm's method

1NN-lbp-G	0.005
1NN-lbp-L	0.005556
1NN-pca-G	0.00625
1NN-grey-G	0.007143
PCALDA-L	0.008333
SVM-lbp-G	0.01
SVM-pca-G	0.0125
1NN-pca-L	0.016667
PCALDA-G	0.025
1NN-grey-L	0.05

SVM-grey-G

Wilcoxon's Test

	1	2	3	4	5	6	7	8	9	10	11
1NN-grey-G (1)	-				○						
1NN-pca-G (2)		-			○						
1NN-lbp-G (3)			-	○	○	○	○				○
PCALDA-G (4)			●	-						●	
SVM-grey-G (5)	●		●		-					●	
SVM-pca-G (6)						-				●	
SVM-lbp-G (7)							-				
1NN-grey-L (8)								-		●	
1NN-pca-L (9)									-	●	
1NN-lbp-L (10)					○					-	
PCALDA-L (11)											-

Figure 5: Statistical tests performed using the CCR of only cross-database experiments with non-distorted faces. Holm's results with a 95% significance level, Wilcoxon's above the main diagonal 90%, and below it 95%.

not and vice versa, local 1-NNs are superior to the rest in statistical terms. However, when the same type of faces (neutral or distorted) are found in training and test sets, global SVMs and local 1-NNs behave in a statistically equal manner. Note that, under different acquisition conditions for training and test images, that is, in cross-database experiments, none of the classification models showed a statistical superiority over the rest, even though when face images were not distorted.

A clear advantage of one type of feature over the other was not detected; nonetheless the best choice would be to lean towards grey levels or PCA features since it is found more times among the best classification models (considering both measures).

Regarding how the performances of the models are affected by the data sets used, no major differences where detected in most of the cases. However, when using FERET and PAL databases (in both combinations of training with one and testing with the other) lower performances were obtained. This is probably due to the different acquisition conditions of the face images in each database.

There is a considerable variability among the actual CCR results (see Appendix A) which is due to the different levels of difficulty in solving the problem of gender classification of the experiments. Considering only the best CCR per experiment, the lowest and highest CCR's are 72.30% and 96.00%, respectively. Although not all our results can be directly compared to those published in the literature, Bekios-Calfa et al. [23] studied several linear discriminant techniques for gender classification, presenting some experiments that are somehow equivalent to some of those in this work. Table 3 provides the CCR's obtained in each case for a quick comparison, showing that our results are at the same level of the state of the art in face gender classification.

	Single-database Exp.		Cross-database Exp.	
	FERET	PAL	FERET/PAL	PAL/FERET
Bekios-Calfa et al.	93.95%	89.81%	71.50%	78.85%
Our results	94.06%	88.57%	72.30%	77.11%

Table 3: Correct Classification Rates of four experiments compared with the equivalent experiments presented in [23].

6. Conclusion

This paper has presented a comprehensive experimental study on gender classification techniques using non-distorted and distorted faces. An extensive comparison of two representation approaches (local and global), three types of features (grey levels, PCA and LBP) and three classifiers (1-NN, SVM and PCA+LDA) has been provided by means of three statistical tests applied to two performance measures (CCR and d').

From the findings of these statistical tests, we can see that global as well as local approaches can successfully solve gender classification problems when the conditions present in the test images can also be found in the training set.

According to the results of both performance measures, in the case of training and test sets with different face distortions, local approaches significantly outperform global solutions. This superiority is due to the fact that local approaches are designed to better deal with distortions in the face images and an expressive/occluded face can be seen as a distorted neutral face and vice versa.

In cross-database experiments with neutral faces, global classification models do not statistically surpass local models. Moreover, some classification models are never found among the models that achieve the best performances for each group of experiments; this is the case of global 1-NNs.

To conclude, global SVMs together with local 1-NNs are the best classification models to address gender classification problems among those considered in this work. However, when training and test images do not share the same type of distortions or acquisition conditions, local 1-NN approaches surpass global solutions. Focusing on the three different types of features analysed, no statistical differences were found among them.

7. Acknowledgements

This work has been partially funded by Universitat Jaume I through grant FPI PREDOC/2009/20 and project P1-1B2012-22, and project TIN2009-14205-C04-04 from the Spanish Ministerio de Economía y Competitividad.

Appendix A. Correct Classification Rates

The CCR obtained for each one of the experiments carried out are shown in Table A.4.

Appendix B. $D-$prime Measure

The d' values obtained for each one of the experiments carried out are shown in Table B.5.

References

[1] A. M. Martinez, Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class, IEEE Trans Pattern Anal Mach Intell 24 (6) (2002) 748–763.

[2] L. Chen, H. Man, A. Nefian, Face recognition based on multi-class mapping of fisher scores, Pattern Recognition 38 (6) (2005) 799–811.

[3] T. Ahonen, A. Hadid, M. Pietikäinen, Face description with local binary patterns: application to face recognition, IEEE Trans Pattern Anal Mach Intell 28 (12) (2006) 2037–2041.

[4] A. N. Rajagopalan, K. Rao, Y. Kumar, Face recognition using multiple facial features, Pattern Recognit Lett 28 (3) (2007) 335–341.

[5] H. Wang, X. Wu, Eigenblock approach for face recognition, Int J Comput Intell Research 3 (1) (2007) 72–77.

[6] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, Y. Ma, Robust face recognition via sparse representation, IEEE Trans Pattern Anal Mach Intell 31 (2) (2009) 210–227.

[7] M. Lyons, J. Budynek, S. Akamatsu, Automatic classification of single facial images, IEEE Trans Pattern Anal Mach Intell 21 (12) (1999) 1357–1362.

[8] G. Donato, M. S. Bartlett, J. C. Hager, P. Ekman, T. J. Sejnowski, Classifying facial actions, IEEE Trans Pattern Anal Mach Intell 21 (10) (1999) 974–989.

[9] B. Fasel, J. Luettin, Automatic facial expression analysis: a survey, Pattern Recognition 36 (1) (2003) 259–275.

[10] F. De la Torre, J. F. Cohn, Guide to Visual Analysis of Humans: Looking at People, Springer, 2011, Ch. Facial Expression Analysis, pp. 1–31.

[11] E. Makinen, R. Raisamo, Evaluation of gender classification methods with automatically detected and aligned faces, IEEE Trans Pattern Anal Mach Intell 30 (3) (2008) 541 –547.

[12] B. Moghaddam, Learning gender with support faces, IEEE Trans Pattern Anal Mach Intell 24 (5) (2002) 707–711.

[13] W. Zhao, R. Chellappa, Face Processing: Advanced Modeling and Methods, Academic Press, 2006.

[14] A. J. O'Toole, K. A. Deffenbacher, D. Valentin, K. McKee, D. Huff, H. Abdi, The perception of face gender: the role of stimulus structure in recognition and classification, Memory & cognition 26 (1) (1998) 146–160.

[15] J. Y. Baudouin, G. Tiberghien, Gender is a dimension of face recognition, J Exp Psychol Learn 28 (2) (2002) 362–365.

[16] G. Hole, V. Bourne, Face Processing. Psychological, Neuropsychological, and Applied Perpectives, Oxford University Press, 2010.

[17] S. Buchala, N. Davey, R. J. Frank, M. Loomes, T. M. Gale, The role of global and feature based information in gender classification of faces: a comparison of human performance and computational models, Int J Neural Syst 15 (1-2) (2005) 121–128.

[18] W. Yang, C. Chen, K. Ricanek, C. Sun, Gender classification via global-local features fusion, LNCS 7098 (2011) 214–220.

[19] M. Toews, T. Arbel, Detection, localization, and sex classification of faces from arbitrary viewpoints and under occlusion, IEEE Trans Pattern Anal Mach Intell 31 (9) (2009) 1567–1581.

[20] T. Jabid, M. H. Kabir, O. Chae, Gender classification using local directional pattern (ldp), ICPR (2010) 216–2165.

[21] C. Shan, Learning local binary patterns for gender classification on real-world face images, Pattern Recognit Lett 33 (4) (2012) 346–349.

[22] M. Nazir, E. Sciences, S. Arabia, Multi-view gender classification using hybrid transformed features, Int J Mult Ubiqu Eng 7 (2) (2012) 515–520.

[23] J. Bekios-Calfa, J. M. Buenaposada, L. Baumela, Revisiting linear discriminant techniques in gender recognition, IEEE Trans Pattern Anal Mach Intell 33 (4) (2011) 585–564.

[24] P. Viola, M. Jones, Robust real-time face detection, Int J Comp Vis 57 (2004) 137–154.

[25] G. R. Bradski, A. Kaehler, Learning OpenCV, O'Reilly, 2008.

[26] K. Turkowski, Filters for common resampling tasks, Graphics Gems I (1990) 147–165.

[27] K. Fukunaga, Introduction to Statistical Pattern Recognition, Academic Press, 1990.

[28] H. Oh, K. Lee, S. Lee, Occlusion invariant face recognition using selective local non-negative matrix factorization basis images, Image and Vision Computing 26 (11) (2008) 1515–1523.

[29] S. Kumari, P. K. Sa, B. Majhi, Gender classification by principal component analysis and support vector machine, Proc Int Conf Communication, Computing and Security - ICCCS (2011) 339–342.

[30] T. Ojala, M. Pietikainen, T. Maenpaa, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Trans Pattern Anal Mach Intell 24 (7) (2002) 971–987.

[31] T. Ojala, M. Pietikäinen, T. Mäenpää, Gray scale and rotation invariant texture classification with local binary patterns, European Conf. Computer Vision 1842 (2000) 404–420.

[32] P. N. Belhumeur, J. P. Hespanha, D. J. Kriegman, Eigenfaces vs. fisherfaces: recognition using class specific linear projection, IEEE Trans Pattern Anal Mach Intell 19 (7) (1997) 711–720.

[33] C. Cortes, V. Vapnik, Support-vector networks, Machine Learning 20 (1995) 273–297.

[34] C.-C. Chang, C.-J. Lin, Libsvm: A library for support vector machines, ACM Trans Intell Syst Technol 2 (3) (2011) 27:1–27:27.

[35] P. J. Phillips, S. A. Rizvi, P. J. Rauss, The FERET evaluation methodology for face-recognition algorithms, IEEE Trans Pattern Anal Mach Intell 22 (10) (2000) 1090–1104.

[36] M. Minear, D. Park, A lifespan database of adult facial stimuli, Behav Res Meth Instr C 36 (2004) 630–633.

[37] A. Martinez, R. Benavente, The AR face database, Tech. rep., CVC Technical Report no.24 (1998).

[38] J. Alcalá-Fdez, A. Fernandez, J. Luengo, J. Derrac, S. García, L. Sánchez, F. Herrera, KEEL data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework, Journal of Multiple-Valued Logic and Soft Computing 17 (2-3) (2011) 255–287.

[39] R. Iman, J. Davenport, Approximations of the critical region of the friedman statistic, Communications in Statistics 9 (1980) 571–595.

[40] H. S., A simple sequentially rejective multiple test procedure, Scandinavian Journal of Statistics 6 (1979) 65–70.

[41] F. Wilcoxon, Individual comparisons by ranking methods, Biometrics Bulletin 1 (6) (1945) 80–83.

[42] C. A. Bellera, M. Julien, J. A. Hanley, Normal approximations to the distributions of the wilcoxon statistics: Accurate to what n? graphical insights, Journal of Statistics Education 18 (2) (2010) 1–17.

[43] J. Derrac, S. García, D. Molina, F. Herrera, A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms, Swarm and Evolutionary Computation 1 (1) (2011) 3–18.

Training Data Set	Test Data Set	Global							Local			
		NN			PCA+LDA	SVM			NN			PCA+LDA
		Grey Levels	PCA	LBP		Grey Levels	PCA	LBP	Grey Levels	PCA	LBP	
FERET 58:42	FERET	85.31	85.57	86.40	91.86	93.66	92.83	94.06	92.35	91.29	85.58	85.07
	PAL	66.03	64.98	58.19	71.25	66.72	62.55	71.60	66.03	62.19	43.03	60.80
	AR Neutral	79.17	82.31	75.37	77.69	81.54	84.62	84.69	86.15	86.92	61.20	83.08
	AR Light Dis.	82.79	82.60	70.21	78.39	81.07	82.60	83.80	86.99	86.62	64.71	79.73
	AR Heavy Dis.	76.06	74.90	67.43	72.84	74.00	76.71	78.74	83.66	83.14	63.12	76.32
PAL 39:61	FERET	66.53	65.56	71.49	75.22	72.99	70.66	69.62	63.16	62.07	56.50	77.11
	PAL	77.42	77.35	79.23	82.72	85.23	85.61	88.57	83.73	83.52	79.06	73.69
	AR Neutral	81.25	82.31	85.07	89.23	92.31	91.54	88.63	90.00	90.00	88.81	87.69
	AR Light Dis.	80.88	80.69	80.27	82.03	85.66	84.32	85.14	86.99	85.66	81.21	67.88
	AR Heavy Dis.	75.68	75.80	76.30	74.00	77.48	75.93	79.73	78.12	76.96	75.92	65.51
AR Neutral 57:43	FERET	76.02	76.86	65.23	80.09	80.83	77.21	65.29	78.90	78.90	69.86	78.20
	PAL	73.35	72.30	68.12	71.43	75.09	70.38	72.13	74.39	73.17	74.56	65.51
	AR Neutral	83.99	82.46	88.81	87.54	90.42	98.15	91.96	88.92	89.08	92.24	86.31
	AR Light Dis.	88.18	87.76	85.28	85.66	88.30	94.65	88.96	89.79	89.45	87.29	85.32
	AR Heavy Dis.	82.47	82.34	82.10	80.46	82.52	92.66	83.27	85.95	85.69	84.36	83.53
AR Light Distortions 57:43	FERET	72.59	72.94	69.20	76.56	77.66	75.22	74.52	80.59	81.23	75.72	77.41
	PAL	72.47	72.65	69.51	72.64	76.48	73.52	73.39	73.69	73.34	74.74	65.85
	AR Neutral	91.23	91.38	90.00	91.08	95.93	96.92	93.13	95.54	94.62	92.84	86.15
	AR Light Dis.	91.24	91.24	87.70	92.82	95.66	99.07	92.17	94.22	93.69	91.20	86.42
	AR Heavy Dis.	85.15	85.22	83.80	85.28	88.89	92.79	86.61	89.32	88.65	87.48	83.14
AR Heavy Distortions 57:43	FERET	71.50	72.10	68.65	73.58	75.77	75.22	69.22	82.17	82.97	74.53	77.56
	PAL	72.82	72.82	69.34	70.56	72.82	70.38	72.69	74.91	71.43	73.87	66.03
	AR Neutral	90.46	90.31	90.45	90.46	94.78	98.46	92.66	96.00	94.92	92.39	85.69
	AR Light Dis.	91.05	90.90	88.31	91.43	95.73	95.41	91.48	94.23	93.46	91.46	85.85
	AR Heavy Dis.	87.57	87.28	85.83	89.21	91.85	98.06	89.72	91.02	90.60	89.02	83.14

Table A.4: Correct classification rates (%) obtained in all experiments. Class balances (percentage of male:female faces) are shown below each training data set.

Training Data Set	Test Data Set	Global							Local			
		NN			PCA+LDA	SVM			NN			PCA+LDA
		Grey Levels	PCA	LBP		Grey Levels	PCA	LBP	Grey Levels	PCA	LBP	
FERET 58:42	FERET	2.08	2.11	2.23	2.78	3.04	2.92	3.12	3.13	3.18	2.69	2.16
	PAL	1.14	1.11	1.19	1.44	1.66	1.38	1.53	1.77	1.67	1.13	1.01
	AR Neutral	1.96	1.88	2.11	1.97	2.19	2.39	2.44	2.36	2.55	1.25	2.29
	AR Light Dis.	1.89	1.88	1.27	1.74	1.98	1.99	2.24	2.42	2.45	1.98	2.13
	AR Heavy Dis.	1.39	1.32	0.98	1.18	1.26	1.43	1.71	2.04	2.08	1.79	1.78
PAL 39:61	FERET	1.26	1.22	1.16	1.51	1.82	2.05	1.39	2.13	1.86	1.94	1.47
	PAL	1.46	1.46	1.74	1.86	2.05	2.08	2.37	2.21	2.19	2.27	1.19
	AR Neutral	1.89	1.83	2.44	2.39	2.89	2.74	2.88	2.55	2.58	3.28	2.35
	AR Light Dis.	1.76	1.74	1.73	1.82	2.12	2.05	2.32	2.42	2.36	3.07	1.98
	AR Heavy Dis.	1.41	1.42	1.46	1.28	1.51	1.48	1.92	2.03	2.01	2.58	1.68
AR Neutral 57:43	FERET	1.52	1.55	1.05	1.76	1.77	1.57	1.21	1.57	1.57	1.07	1.52
	PAL	1.07	1.12	0.83	1.06	1.33	1.08	1.29	1.27	1.22	1.58	0.84
	AR Neutral	1.98	1.86	2.46	2.34	2.63	2.46	2.84	2.52	2.53	3.31	2.37
	AR Light Dis.	2.37	2.34	2.09	2.18	2.49	2.47	2.46	2.66	2.64	2.63	2.22
	AR Heavy Dis.	1.88	1.88	1.82	1.74	1.95	1.89	1.93	2.30	2.30	2.50	2.02
AR Light Distortions 57:43	FERET	1.32	1.34	1.21	1.44	1.51	1.53	1.36	1.69	1.75	1.43	1.48
	PAL	1.13	1.15	0.92	1.39	1.40	1.19	1.19	1.24	1.24	1.67	0.83
	AR Neutral	2.82	2.99	2.57	2.71	3.44	3.21	3.01	3.59	3.50	3.42	2.29
	AR Light Dis.	2.74	2.76	2.33	2.84	3.48	3.31	2.84	3.35	3.28	3.08	2.36
	AR Heavy Dis.	2.18	2.19	1.96	2.08	2.49	2.47	2.21	2.76	2.72	2.86	1.96
AR Heavy Distortions 57:43	FERET	1.27	1.29	1.16	1.25	1.39	1.53	1.09	1.81	1.89	1.34	1.49
	PAL	1.15	1.16	0.91	1.15	1.13	0.98	1.15	1.36	1.17	1.54	0.89
	AR Neutral	2.66	2.71	2.64	2.64	3.21	3.06	2.91	3.73	3.59	3.38	2.31
	AR Light Dis.	2.71	2.69	2.39	2.72	3.44	3.17	2.74	3.37	3.26	3.05	2.29
	AR Heavy Dis.	2.29	2.26	2.14	2.46	2.77	2.68	2.53	2.77	2.73	2.89	1.99

Table B.5: D-prime values for all experiments. Class balances (percentage of male:female faces) are shown below each training data set.