Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2009, Article ID 184617, 19 pages
doi:10.1155/2009/184617

Research Article

Recognition of Faces in Unconstrained Environments: A Comparative Study

Javier Ruiz-del-Solar, Rodrigo Verschae, and Mauricio Correa

Department of Electrical Engineering, Universidad de Chile, Avenida Tupper 2007, 837-0451 Santiago, Chile

Correspondence should be addressed to Javier Ruiz-del-Solar, jruizd@cec.uchile.cl

Received 10 October 2008; Revised 31 January 2009; Accepted 13 March 2009

Recommended by Kevin Bowyer

The aim of this work is to carry out a comparative study of face recognition methods that are suitable to work in unconstrained environments. The analyzed methods are selected by considering their performance in former comparative studies, in addition to be real-time, to require just one image per person, and to be fully online. In the study two local-matching methods, histograms of LBP features and Gabor Jet descriptors, one holistic method, generalized PCA, and two image-matching methods, SIFT-based and ERCF-based, are analyzed. The methods are compared using the FERET, LFW, UCHFaceHRI, and FRGC databases, which allows evaluating them in real-world conditions that include variations in scale, pose, lighting, focus, resolution, facial expression, accessories, makeup, occlusions, background and photographic quality. Main conclusions of this study are: there is a large dependence of the methods on the amount of face and background information that is included in the face's images, and the performance of all methods decreases largely with outdoor-illumination. The analyzed methods are robust to inaccurate alignment, face occlusions, and variations in expressions, to a large degree. LBP-based methods are an excellent election if we need real-time operation as well as high recognition rates.

1. Introduction

Many different face-recognition approaches have been developed in the last few years [1–4], ranging from classical Eigenspace-based methods (see, e.g., eigenfaces [5]), to sophisticated systems based on thermal's information, high-resolution images, or 3D models (see, e.g., [4, 6, 7]). However, the recognition of faces in unconstrained environments has not been completely solved [8]. In addition, some time-demanding applications, such as searching faces in nonannotated or partially annotated databases (i.e., news databases, the Internet, etc.) and HRI (Human-Robot Interaction), impose extra requirements of real-time operation, just one image per person and fully on-line operation (no off-line enrollment), which are difficult to achieve.

In this general context, the aim of this article is to carry out a comparative study of face-recognition methods by considering these requirements. The main motivation is the lack of direct and detailed comparisons of this kind of methods under the same conditions. The results of this comparative study are a guide for developers of face-recognition systems. As aforementioned, we concentrate ourselves on methods that fulfill the following requirements: (i) *full on-line operation*: no off-line enrollment stages. All processes must run on-line. The system has to be able to build the face database incrementally from scratch; (ii) *real-time operation*: the recognition process should be fast enough to allow real-time interaction in case of HRI or to search large databases in reasonable time (a few seconds or a couple of minutes depending on the application and the size of the database); (iii) *one single image per person problem*: one two-dimensional face image of an individual should be enough for his/her later identification. Databases containing just one face image per person should be considered. The main reasons are savings in storage and computational costs and the impossibility of obtaining more than one face image from a given individual in certain situations. In addition, we want to consider standard 2D images, and not high-resolution, 3D or thermal images that are not always available and that can slow down the recognition process; (iv) *unconstrained*

environments: no restrictions over environmental conditions such as scale, pose, lighting, focus, resolution, facial expression, accessories, makeup, occlusions, background, and photographic quality are required.

Thus, in this study two local-matching, one holistic, and two novel image-matching methods are selected by considering their fulfillment of the aforementioned requirements and their performance in former comparative studies of face-recognition methods [2, 9–12]. The two local-matching methods, namely, histograms of LBP (Local Binary Patterns) features [13] and Gabor-Jet features with Borda count classifiers [10] are selected considering their performance in the studies reported in [2, 10]. Among the holistic methods, a member of the eigenspace-based family of face-recognition methods is included, generalized PCA (Principal Component Analysis) with Euclidian distance and modified LBP features to achieve illumination invariance [11] (the restriction of *one single image per person* does not allow to include easily other members of the family). In addition, two novel face-recognition methods based on advanced image-matching methods are also considered: SIFT (Scale-Invariant Feature Transform) descriptors with local and global matching methods [12] and ERCF (Extremely Randomized Clustering Forest) of SIFT Descriptors used together with linear classifiers [14]. This last method, although not being real-time, is included for comparison purposes, because of the excellent results it has obtained in the LFW database [15].

The comparative study is carried out using the FERET [10], LFW (Labeled Faces in the Wild) [8], UCHFaceHRI [12], and FRGC (Face Recognition Grand Challenge) databases [16, 17]. We choose to use the very well-known FERET database, because it is one of the most employed face databases, and therefore it allows comparing results to other studies. In addition, we think that robustness when using a large database is also important and FERET contains more than 1,000 individuals. We include the LFW database because it is specially designed to study the problem of unconstrained face recognition. It corresponds to a set of more than 13,000 images of faces collected from the web, images which exhibit natural variability in pose, lighting, focus, resolution, facial expression, age, gender, race, accessories, make-up, occlusions, background, and photographic quality. The only constraint on these faces is that they were detected using the Viola-Jones face detector [18]; therefore, they correspond to frontal and quasifrontal faces. We also include in this study the new UCHFaceHRI, which is especially designed to compare face analysis methods for HRI. This database contains 30 individuals and includes images with natural variations in illumination (indoor and outdoor), scale, pose, and expressions. Finally, we consider experiments using the FRGC dataset, whose data corpus consists of 50,000 recordings, divided into training and validation partitions. We used FRGC's experiments 1 and 4, designed to measure progress on recognition from controlled and uncontrolled frontal face images. Thus, the comparative study includes 4 stages. (1) In the first stage all methods (except ERCF) are compared using the FERET database. Aspects such as variable illumination, alignment's accuracy, occlusions, and dependence on the database's size are measured, and the results are analyzed in terms of recognition rate and computational costs. (2) Some selected methods are further analyzed using the more challenging conditions defined by LFW. In addition to all the variability expressed in the LFW images, we analyze the dependence of the methods on the alignment's accuracy as well as on the amount of background and face's information considered in the analysis of the images. (3) The best variants of each of these methods (including selected distance's metrics and cropping-size for each case) are further analyzed using the natural requirements defined in the UCHFaceHRI database. (4) Finally, the best performing methods in all tests are analyzed and compared to state-of-the-art methods using the FRGC database. This study corresponds to an extended version of the one presented in [19].

This paper is structured as follows. The methods under analysis are described in Section 2. In Sections 3–6 the comparative analysis of these methods is presented. Finally, in Section 7 results are discussed, and conclusions are given.

2. Methods under Comparison

As mentioned above, the algorithms' selection criteria are their fulfillment of the defined requirements, and their performance in former comparative studies of face-recognition methods [2, 9–12]. In the comparison we decided to consider local-matching, holistic, and advanced image-matching methods.

Local-matching methods behave well when just one image per person is available [2], and some of them have presented very good results in standard databases such as FERET [10]. Thus, taking into account the results of [10], and our requirements of high-speed operation, we selected two methods to be analyzed. The first one is based on the use of histograms of LBP features, and the second one is based on the use of Gabor filters and Borda count classifiers.

When analyzing which holistic methods to include, the first idea was to consider methods based on eigenspace-decompositions (see a basic categorization in [9]). However, these methods normally fail when just one image per person is available, mainly because they have difficulties to build the required representation models. This difficulty can be overcome if a generalized face representation is built. Such representation can be built using a generalized PCA model. Thus, we incorporated to the study a face-recognition method based on a generalized PCA model.

We also decided to consider in this study advanced image-matching methods, which are not very popular in the face-recognition community, but which have been successfully applied in other computer vision contexts. Thus, taking into account that local interest points and descriptors (see, e.g., SIFT [20]) have been already used to solve successfully some other biometric problems (see, e.g., fingerprint verification [21] and off-line signature verification [22]), and as a first stage of complex face-recognition systems [6], we decided to test the suitability of a SIFT-based face-recognition system in this study. Finally, we also included

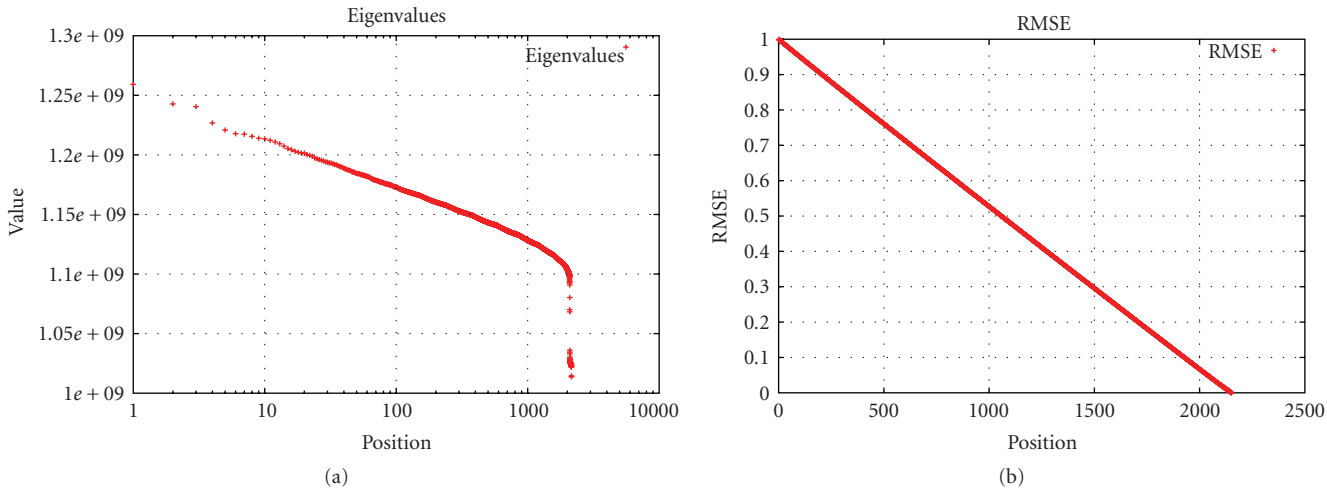


Figure 1: (a) Spectrum of the eigenvalues in the employed generalized PCA representation. Training set of size 2,152. (b) RMSE of the employed representation.

the recently proposed ERCF [14], a tree-based classification method designed to verify if a pair of images corresponds to the same object or not. The reason to include this last method in the comparison is the excellent results that have been obtained in recognizing faces in the LFW database [15]. The use of SIFT features for face authentication was investigated in [23] for the first time; however, no comparisons with other methods were presented.

The aforementioned selected methods are described in the next sections.

2.1. Generalized PCA. We implemented a face-recognition method that uses generalized PCA as projection algorithm, the Euclidian distance as similarity measure, and modified LBP features [24]. We used a generalized PCA approach, which consists on building a PCA representation in which the model does not depend on the individuals to be included in the final database, that is, on their face's images, because the PCA projection model is built using face's images that belong to a different set of persons. This allows applying this method in the case when just one single image per person is available. Our PCA model was built using 2,152 face images obtained from different face databases and the Internet. For compatibility with the results presented in [9], the model was built using face images scaled and cropped to 100×185 pixels and was aligned using eye's information. Using a similar approach to the one described in [16], we analyzed the validity of this generalized PCA representation by verifying that the main part of the eigenspectrum, that is, the spectrum of the ordered eigenvalues, is approximately linear between the 10th and 1,500th components, using a logarithmic scale for the components (see Figure 1(a)). The RMSE [9] was used as a criterion to select the appropriate number of components to be used. To achieve a RMSE between 0.9 and 0.5, the number of employed PCA components has to be in the range of 200 to 1,050 (see Figure 1(b)). Taking into account these results as well as the tradeoff between number of components and speed, we choose to implement two flavors of our system, one with 200 components and one with 500. Modified LBP features were used because according to the study presented in [11], this feature-space transformation (together with SQI) is one of the most suitable algorithms to achieve illumination compensation and normalization in eigenspace-based face-recognition systems.

2.2. LBP Histograms. Face recognition using histograms of LBP features was originally proposed in [13] and used by many groups since then. In the original approach, three different levels of locality are defined: pixel level, regional level, and holistic level. The first two levels of locality are realized by dividing the face image into small regions from which LBP features are extracted and histograms are used for efficient texture information representation. The holistic level of locality, that is, the global description of the face, is obtained by concatenating the regional LBP extracted features. The recognition is performed using a nearest neighbor classifier in the computed feature space using one of the three following similarity measures: histogram intersection, log-likelihood statistic, and Chi square. We implemented this recognition system, without considering preprocessing (cropping using an elliptical mask and histogram equalization are used in [13]), and by choosing the following parameters: (i) images divided in 10 (2×5), 40 (4×10), or 80 (4×20) regions, instead of using the original divisions which range from 16 (4×4) to 256 (16×16), and (ii) the mean square error as similarity measure, instead of the log-likelihood statistic. We also carried out preliminary experiments for replacing the LBP features by modified LBP features, but better results were always obtained by using the original LBP features. Thus, considering the 3 different image divisions and the 3 different similarity measures, we get 9 flavors of this face-recognition method.

2.3. Gabor Jets Descriptors. Local-matching approaches for face recognition are compared in [10]. The study analyzes

several local feature representations, classification methods, and combinations of classifier alternatives. Taking into account the results of their study, the authors implemented a system that integrates the best possible choice at each step. That system uses Gabor jets descriptors as local features, which are uniformly distributed over the images, one wavelength apart. In each grid position of the test and gallery image and at each scale (multiscale analysis), the Gabor jets are compared using normalized inner products, and these results are combined using the Borda count method. In the Gabor feature representation, only Gabor magnitudes are used, and 5 scales and 8 orientations of the Gabor filters are adopted. We implemented this system using all parameters described in [10] (filter frequencies and orientations, grid positions, face image size).

2.4. SIFT Descriptors. Wide-baseline matching approaches based on local interest points and descriptors have become increasingly popular and have experienced an impressive development in recent years. Typically, local interest points are extracted independently from both a test and a reference image and then characterized by invariant descriptors, and finally the descriptors are matched until a given transformation between the two images is obtained. Lowe's system [20] using SIFT descriptors and a probabilistic hypothesis rejection-stage is a popular choice for implementing object-recognition systems, given its recognition capabilities, and near real-time operation. However, Lowe's system's main drawback is the large number of false positive detections. This drawback can be overcome by the use of several hypothesis rejection stages as, for example, in the L&R system [21]. This system has already been used in the construction of robust fingerprint verification systems [21] and for off-line signature verification [22]. Here, we use the L&R system to build a face-recognition system, with three different flavors. In the first one, *Full*, all verification stages defined in [21] are used, while in the second one, *Simple*, just the probabilistic hypothesis rejection stages are employed. In the third one, *Matches*, the number of matching key points without using any rejection stages is considered.

2.5. ERCF: Extremely Randomized Clustering Forest. In [14] a robust method to learn a similarity measure is proposed, which allows to discriminate whether a pair of object's images corresponds to the same object or not (the objects could be faces). The method is especially designed to be used in object recognition problems and makes use of ERCF and SIFT descriptors. The learning is done for specific object classes, such as frontal faces or specific views of cars. The method basically consists of three stages. In the first stage, pairs of similar patches, measured in terms of a normalized cross-correlation, are selected. In the second stage, each pair of patches is coded (quantized) by means of an ERCF of SIFT descriptors. ERCF is a sparse representation of the image that is built using classification trees. Each classification tree is generated using SIFT descriptors and used for vector quantization. In the third stage, the quantized pairs of patches are used to build a feature vector, which is finally used to evaluate the similarity of the image pair using a linear classifier. In this study we use the author's implementation of the method, available on http://lear.inrialpes.fr/people/nowak/similarity/index.html.

2.6. Notation: Methods and Variants. We use the following notation to refer to the methods and their variations: A, B, and C. (i) A describes the name of the face-recognition algorithm: H is Histogram of LBP features, PCA is generalized PCA with modified LBP features, GJD is Gabor Jets Descriptors, SD is L&R system with SIFT descriptors, and ERCF is Extremely Randomized Clustering Forest; (ii) B denotes the similarity measure: HI is Histogram Intersection, MSE is Mean square error, XS is Chi square, BC is Borda Count, and EU is Euclidian Distance, except for the case of SD and ERCF, which do not use any explicit distance's measure; (iii) C describes additional parameters: number of divisions in the case of the LBP-based method, number of principal components in the case of PCA, size of the reference-set for the GJD case (see explanation in Section 4), and flavor (*Full*, *Simple,* or *Matches*) in the case of SD.

3. Comparative Study Using the FERET Database

Face images are scaled and cropped to 100×185 pixels and 203×251 (for compatibility with former studies [9, 10]), except for the case of the PCA method in which, for simplicity, just one image size (100×185) was employed (the generalized PCA model depends on the image cropping). In all cases, faces are aligned by centering the eyes in the same relative positions, at a fixed distance between the eyes, which was 62 pixels for the 100×185 size images and 68 pixels for the 203×251 size images. The amount of face information and background contained in the cropped images can be measured using the normalized width (*nw*) and height (*nh*), defined as the image width/height divided by the distance between eyes. This means that the *nw*/*nh* of the analyzed images are 1.6/3.0 for images of 100×185 pixels and 3.0/3.7 for images of 201×253 pixels. To compare the methods we used the FERET evaluation procedure [25], which established a common data set and a common testing protocol for evaluating semiautomated and automated face-recognition algorithms. We used the following sets: (i) *fa* set (1,196 images), used as gallery set (contains frontal faces of 1,196 people); (ii) *fb* set (1,195 images), used as test set 1 (in *fb* subjects were asked for a different facial expression than in *fa*); (iii) *fc* set (194 images,) used as test set 2 (in *fc* pictures were taken under different lighting conditions). In all cases the information about the eyes' position provided by FERET was used for the face alignment.

In addition, we carried out extra experiments by adding noise to the position of the eyes in the *fb* set, and also by adding artificial occlusions in these images. The goal was to test the robustness of the different methods. Finally, we also compared the computational performance of the methods. ERCF was not considered in this first comparison, neither in the FRGC experiments, because the method is not real-time and, to carry out all the experiments, it takes a very

Table 1: FERET fa-fb and fa-fc tests. Top-1 recognition rate. Noise in eye positions and face occlusion is tested in the fa-fb test. OR: Original. OC: Original plus Occlusion. The best results for each condition are presented in bold. Methods that have differenc

Method	100×185						203×251					
	fa-fb					*fa-fc*	*fa-fb*					*fa-fc*
	OR	Noise in eye positions			OC		OR	Noise in eye positions			OC	
		2.5%	5%	10%				2.5%	5%	10%		
H-HI-10	95.6	**95.0**	**91.3**	**81.8**	93.6	12.9	95.1	23.7	22.4	16.4	93.4	50.0
H-MSE-10	95.6	**95.0**	**91.3**	**81.8**	93.6	12.9	95.1	23.7	22.4	16.4	93.4	50.0
H-XS-10	95.7	94.7	**92.3**	**82.2**	78.4	14.9	95.1	41.3	39.4	31.0	86.1	60.8
H-HI-40	**96.5**	**96.0**	89.7	70.9	95.1	57.2	96.5	41.0	39.7	27.5	95.2	85.1
H-MSE-40	**96.5**	**96.0**	89.7	70.9	95.1	57.2	96.5	41.0	39.7	27.5	95.2	85.1
H-XS-40	95.5	93.6	87.0	67.4	92.1	47.4	97.4	76.6	71.4	53.8	95.0	88.1
H-HI-80	**97.2**	**95.6**	90.1	71.5	**96.7**	71.1	96.9	61.1	55.7	40.6	96.6	91.8
H-MSE-H-MSE-80	**97.2**	**95.6**	90.1	71.5	**96.7**	71.1	96.9	61.1	55.7	40.6	96.6	91.8
H-XS-80	96.3	94.1	88.3	68.0	94.4	62.9	97.4	87.8	83.9	64.9	**96.7**	92.8
PCA-MSE-200	73.1	55.9	40.7	16.2	63.6	52.1	—	—	—	—	—	—
PCA-MSE-500	76.1	60.3	42.9	16.0	64.9	57.2	—	—	—	—	—	—
GJD-BC	91.4	89.6	85.0	63.1	74.5	**79.9**	**98.5**	95.0	93.6	73.9	**97.7**	**99.0**
SD-FULL	74.3	75.7	73.5	71.5	67.3	7.7	97.1	**96.2**	**95.7**	**95.3**	95.6	67.5
SD-SIMPLE	73.1	75.3	73.1	71.0	68.6	5.7	**97.5**	**96.7**	**96.4**	**96.2**	95.3	63.9
SD-MATCHES	70.3	70.3	67.6	66.7	58.6	4.7	93.9	93.7	94.6	92.3	90.1	44.0

long time. However, the method is considered in the LFW and UCHFaceHRI experiments.

Original fa-fb Test. Table 1 shows *top-1* RR (Recognition Rate) achieved by the different methods under comparison in the original *fa-fb* test, which corresponds to a test with few variations in the acquisition process (uniform illumination, no occlusions). We use the information of the annotated eyes, without adding any noise. From the experiments the following can be observed.

(i) The results obtained with our own implementation of the methods are consistent with those of other studies. The best H-X-X flavors achieved in the 203×251 face images a similar performance (97.4% versus 97%) than the one reported in the original work [13]. GJD-BC achieved a slightly lower performance (98.5% versus 99.5%) than in the original work [10]. When comparing these results to the ones obtained by other authors using more complex systems based on hybrid Gabor-LBP [26], Gabor-Fisher [27], or Fisher-Gabor-LBP [28]—98%, 99% and 99.6%, respectively, we observe that those results are similar or slightly better than ours; however, our systems are much simpler. There are no reports of the use of the generalized PCA or SIFT methods in these datasets.

(ii) The best results ($\sim$ 98.5%) are obtained by GJD-BC, followed by the SD and H-X-80 variants, all using 203×251 images. Nevertheless, other H-X-X variants also get very good results. Interestingly, some H-X-X variants get $\sim$ 97% even using 100×185 size images. The results obtained by the PCA methods are the lowest.

(iii) The performance of the GJD-X-X and SD-X methods depends largely on the normalized size of the cropped images, probably because the methods use information about face shape and contour, which does not appear in the 100×185 images.

Eye Detection Accuracy. Most of the face-recognition methods are very sensitive to face alignment, which depends directly on the accuracy of the eye detection process; eye position is usually the primary, and sometimes the only, source of information for face alignment. For analyzing the sensitivity of the different methods on the eye position's accuracy, we added white noise to the position of the annotated eyes in the *fb* images (see example in Figure 2(a)). The noise was added independently to the x and y eye positions. Table 1 shows the *top-1* RR achieved by the different methods. Our main conclusions are the following.

(i) SD-X methods are almost invariant to the position of the eyes in the case of using 203×251 face images. With 10% error in the position of the eyes, the top-1 RR decreases in just $\sim$2%. The invariance is due to the fact that this method aligns test and gallery images by itself.

(ii) In all other cases the performance of the methods decreases largely with the error in the eye position, probably because they are based on the matching between holistic or feature-based representations of the images. However, if the eye position error is bounded to 5%, the results obtained by some H-X-X variants using 100×185 face images ($\sim$90%) are still acceptable.

Partial Face Occlusions. To analyze the behavior of the different methods in response to partial occlusions on the face area, *fb* face images were divided into 10 different areas (2 columns and 5 rows). One of these areas was randomly selected and its pixels set to 0 (black). See example in Figure 2(b). Thus, in this test each face image of *fb* has one tenth of its area occluded. Table 1 shows the *top-1* RR achieved by the different methods. The main conclusions are as follows.

(i) GJD-BC and H-XS-80 achieve the highest top-1 RR in the 203×251 case, 97.7% and 96.7%, respectively.

(ii) Some H-X-X variants are very robust to face occlusions (e.g., H-HI-10, H-MSE-10, H-X-80) independently of using face images of 100×185 or 203×251 pixels.

(iii) SD-X variants are also robust to occlusions in the 203×251 case.

(iv) PCA is not robust to occlusions; its performance decreases in about 10% compared to the nonoccluded case.

Variable Illumination. Variable illumination is one of the factors with strong influence in the performance of face-recognition methods. Although there are some specialized face databases for testing algorithm invariance against variable illumination (e.g., PIE, YaleB), we choose to use the *fa-fc* test set, because (i) it considers a large number of individuals (394 versus 10 in Yale B and 68 in PIE), and (ii) the illumination conditions are more natural in the *fc* images. Table 1 shows the *top-1* RR achieved by the different methods in this test. The main conclusions are as follows.

(i) The results obtained with our own implementation of the methods are consistent with those of other studies. The best H-X-X flavors achieve in the 203×251 case a higher performance (92.8% versus 79%) than the one reported in the original work [13], probably due to the different image's partitions that we use in our implementation. The best GJD-BC flavors achieve a slightly lower performance (99% versus 99.5%) than the original implementation [10]. When comparing these results to the ones obtained by other authors using more complex systems based on hybrid Gabor-LBP [26], Gabor-Fisher [27], or Fisher-Gabor-LBP [28]—98%, 97% and 99%, respectively—we observe that those results are similar to ours; however, our systems are much simpler. There are no reports of the use of the generalized PCA or SIFT methods in the same database.

(ii) Best performance is achieved by GJD-BC (99%), and second best by H-XS-80 (~93%). In both cases using images cropped to 203×251 pixels.

(iii) In all cases much better results were obtained using larger face images (203×251).

(iv) PCA-X-X and SD-X methods show a low performance in this dataset.

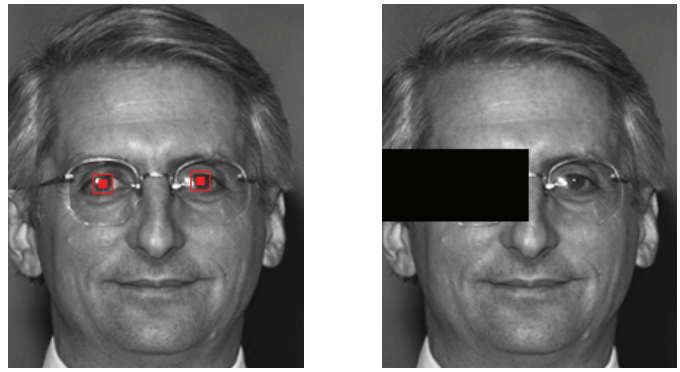
(v) H-X-X methods with a large number of partitions show better performance than variants with a small number of partitions (~93% versus ~50% in the case of using 203×251 images and ~71% versus ~13% in the case of 100×185 images).

Computational Performance. As aforementioned one of the requirements imposed to the methods under comparison is real-time operation. In addition, the memory required by the different methods is very important in some applications where memory could be an expensive resource. Table 2 shows the computational and memory costs of the different methods under comparison, when images of 100×185 are considered. For the case of measuring the computational costs, we considered the feature-extraction time (FET) and the matching time (MT). In the case of measuring memory costs, we considered the database memory (DM), which is the required amount of memory to have the whole database (features) in memory, and the model memory (MM), which is the required amount of memory to have the method model, if any, in memory (PCA matrices for the PCA case and filter bank for the Gabor method). We show the results for databases of 1, 10, 100, and 1,000 individuals (face images). If we consider that in many applications the database size is in the range 10–100 persons, the fastest methods are the H-X-X ones. The second fastest methods are the GJD-BC ones. To achieve real-time operation with a database of 100 or fewer elements, all methods are suitable, except PCA-based methods. In databases of 10–100 individuals, H-X-X and GJD-X-X require less than 8 MBytes of memory (they do not need to keep a model in memory). In the case of H-X-X methods, the required memory increases linearly with the number of partitions.

Summary. As a result of all these experiments we decided to further test these methods in more demanding conditions using the LWF and UCHFaceHRI databases. In this stage we discarded the PCA method, because in all tests it turns to be the weakest one, getting always the lowest scores.

4. Comparative Study Using the LFW Database

The LFW database [8] consists of 13,233 images faces of 5,749 different persons, obtained from news images by means of a face detector (Viola-Jones detector [18]). There are no eyes/fiducial point annotations; the faces were just aligned using the output of the face detector. The faces aligned using the funneling algorithm [29] are also available. The images of the LFW database have a very large degree of variability in the face's expression, age, race, background, and illumination conditions (see Figure 3). Also, unlike other databases, the recognition is only to be done by comparing pairs of images, instead of searching for the most similar face in the database. The idea is that the algorithm being evaluated is given a pair of images, and it has to output whether the two images correspond to the same person or not. There are two evaluation settings already defined by the authors of the LFW: the image restricted setting and the image unrestricted setting. The image restricted setting is the



(a) (b)

Figure 2: Face image of 203 × 251 pixels. (a) Image with eye position (red dot) and square showing a 10% error in the eye position. (b) Image with partial occlusion.

Table 2: Computational and memory costs. FET: Feature Extraction Time. MT: Matching Time. PT: Processing Time. DM: Database Memory. MM: Model Memory. TM: Total Memory. Time measures are in milliseconds; memory measures are in Kbytes. DB sizes of 1, 10, 100, and 1,000 faces are considered. An image size of 100 × 185 pixels is considered.

Method	FET	MT	PT (FET + MT)				DM	MM	TM (DM + MM)			
			1	10	100	1000			1	10	100	1000
H-X-10	15	0.11	15	16	26	120	11	0	11	110	1100	11000
H-X-40	15	0.29	15	18	44	305	41	0	41	410	4100	41000
H-X-80	15	0.42	15	19	57	435	80	0	80	800	8000	80000
PCA-MSE-200	170	0.02	170	170	172	190	0,8	137800	137801	137808	137878	138585
PCA-MSE-500	360	0.02	360	360	362	380	2	137800	137802	137820	137996	139757
GJD-BC	50	0.25	50	53	75	300	33	1240	1273	1572	4559	34427
SD-X	4.7	1.03	6	15	108	1036	428	0	428	4284	42845	428451

most difficult one, and it is the one considered here. Under this setting the only information that the algorithm can use is the image pair; no information of the identity of the faces in the images can be used, that is, the algorithm is restricted to work only using the image pair at hand. The systems are trained (if required) and evaluated using a 10-fold validation procedure, where the folds are symmetric in the sense that the number of matching pairs and nonmatching pairs is the same. See [8] for details.

In the first experiments (Sections 4.1 and 4.2) images were cropped to 100 × 185 pixels (see Figures 4(a) and 4(b)). Given that the mean distance between eyes is 42 pixels, the normalized width and height are $nw = 2.4$ and $nh = 4.4$. We analyze and compare two cases, unaligned and aligned. In the unaligned case, face images have a coarse alignment, which is the one produced by the face detector that was used to obtain the images. In the aligned case, the funnelling algorithm is used to obtain a more accurate alignment. Afterwards, in Section 4.3, all methods are analyzed, considering different region sizes, where the face's images are cropped considering larger and smaller bounding boxes. These experiments analyze the effect of using different amounts of background and face's information in the recognition process (see Figure 4).

Given that the LFW database only requires comparing pairs of faces, and that an important part of the GJD method is the ranking done using Borda count, we had to adapt it to this condition. To accomplish this, we first define a reference set of faces, which is built by randomly selecting face images (e.g., 50) of the same characteristics than the ones under comparison. Then, we take one of the two face images under comparison, and we compare it against the images of the reference set plus the second image under comparison. The relative ranking, computed using Borda count, obtained by the second face image is considered as a measure of the similarity between the pair of images. To obtain a symmetric similarity measure, we repeated the same procedure by switching the roles of the two images, and then averaging the two obtained rankings. The average value was taken as the final similarity measure of the pair of images. We considered three different sizes for the reference set: 10, 50, and 100 faces. To show the importance of using Borda count method, results using the Euclidean distance between the GJD descriptors are also given for comparative purposes.

SD-Full does not work properly in this database, and consequently its results are omitted. In addition, when using the LBP-based methods, HI and MSE always obtained the same recognition results, and therefore the HI case is also omitted. The results corresponding to ERCF consider complete images (250 × 250), and they correspond to those

Figure 3: Examples of faces from the LFW, randomly selected from people with name starting with A.

Figure 4: Examples of faces with different cropping (LFW database). (a) 100×185, unaligned; (b) 100×185, aligned; (c) 81×150, aligned; (d) 122×225, aligned; (e) 125×125, aligned; (f) 250×250, aligned. The last row shows the normalized image's width and height (nw/nh). The images are shown maintaining their relative sizes.

presented in [15]. We use the original results although in our own experiments we got very similar results.

4.1. Experiments Using Unaligned Faces. Table 3 (second and third columns) shows the results for all methods under comparison in the unaligned LFW database. It should be remembered that in the unaligned LFW, all images have a coarse alignment. In all cases (except for ERCF), regions of 100×185 pixels containing the centered face in the 250×250 image were cropped ($nw = 2.4$ and $nh = 4.4$). As it can be observed, the results obtained with our own implementation of the methods are consistent with those of other studies results (in terms of the relative order of the classification accuracy). However the accuracies are low, going form 60% to 72%, values that show the difficulty of the database at hand. In the case of the H-X-X methods,

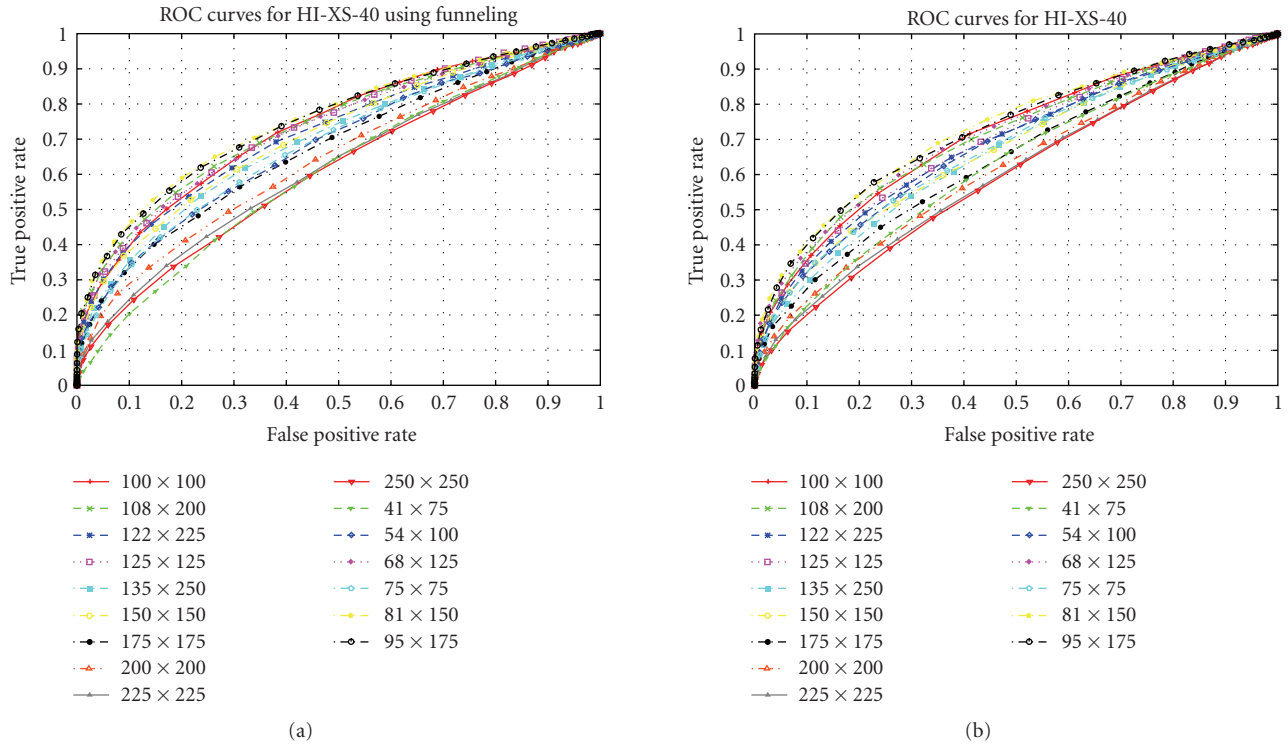


Figure 5: Effect of the image's region size on the performance of the H-XS-40. (a) Faces aligned using funneling; (b) unaligned faces.

best results are obtained with H-X-80, that is, when using the largest number of divisions. The difference between using the Chi-Square and the Mean Square Error is not significant, although the Chi-Square measure gives slightly better results in all cases. For the method based on the GJD, best results are obtained when using the proposed Borda count methodology (it increases the performance in circa 2% over the Euclidean distance); 100 reference images gives slightly better results than 10 or 50. Both methods based on SD got the lowest performance (about 60%–62%). The performance of ERCF is quite good, being ∼ 4% larger than the second best method (GJD-BC-100).

4.2. Experiments Using Aligned Faces. The faces were aligned using the funneling algorithm [29]. Funneling is an unsupervised algorithm for object alignment based on the concept of congealing. Congealing basically consists of searching a sequence of transformations (in this case affine transforms and translations) that are applied to a set of images in order to minimize an entropy measure on the set of images. After having built the congealing model, the transformations can be applied to an unseen image (funneling it) to obtain an aligned image. The main advantage of this method is that it can work in complex objects and that it does not require any labeling during training.

Table 3 (last two columns) shows the results for all methods under comparison using aligned faces. As in the case of unaligned faces (except for ERCF); the face region was cropped considering a region of 100×185 pixels centered in the 250×250 image ($nw = 2.4$ and $nh = 4.4$). Compared to the case of unaligned faces, all methods, but GJD-X-X and SD-Simple, improve or maintain their performance. The H-X-X methods obtain the largest improvement, 2% to 3%, depending on the variant being considered. Again, in the case of LBP based methods, best results are obtained with H-X-80, that is, when using the largest number of divisions, and the Chi-Square distance's measure, with a performance similar to GJD. For the variants based on the GJD, best results are obtained when Borda Count is used (it increases the performance in circa 3% over the Euclidean distance), and 100 reference images gives slightly better results than 10 or 50. However, in this case, the results were slightly worse than the ones obtained for the case of unaligned faces. Again, best results are obtained by ERCF, but this time being about 5% over the second best method.

4.3. Experiments Using Different Windows Sizes. In this section we analyze the effect of using different region sizes in the performance of the analyzed methods. Note that increasing the size of the regions corresponds to adding or removing different amounts of background to the region being analyzed, given that we are not decreasing the scale of the faces. The experiments were performed considering squared image regions, ranging from 50×50 to 250×250, with a step of 25 pixels, and considering regions of ratio $1 : 1.85$ (as in the previous section), ranging from 41×75 to 135×250, with a step of 25 pixels. Results are presented in Figures 5–8 in form of ROC curves. By observing the results, the first thing we can see is the importance of the relative size of the region, that is, the amount of face and

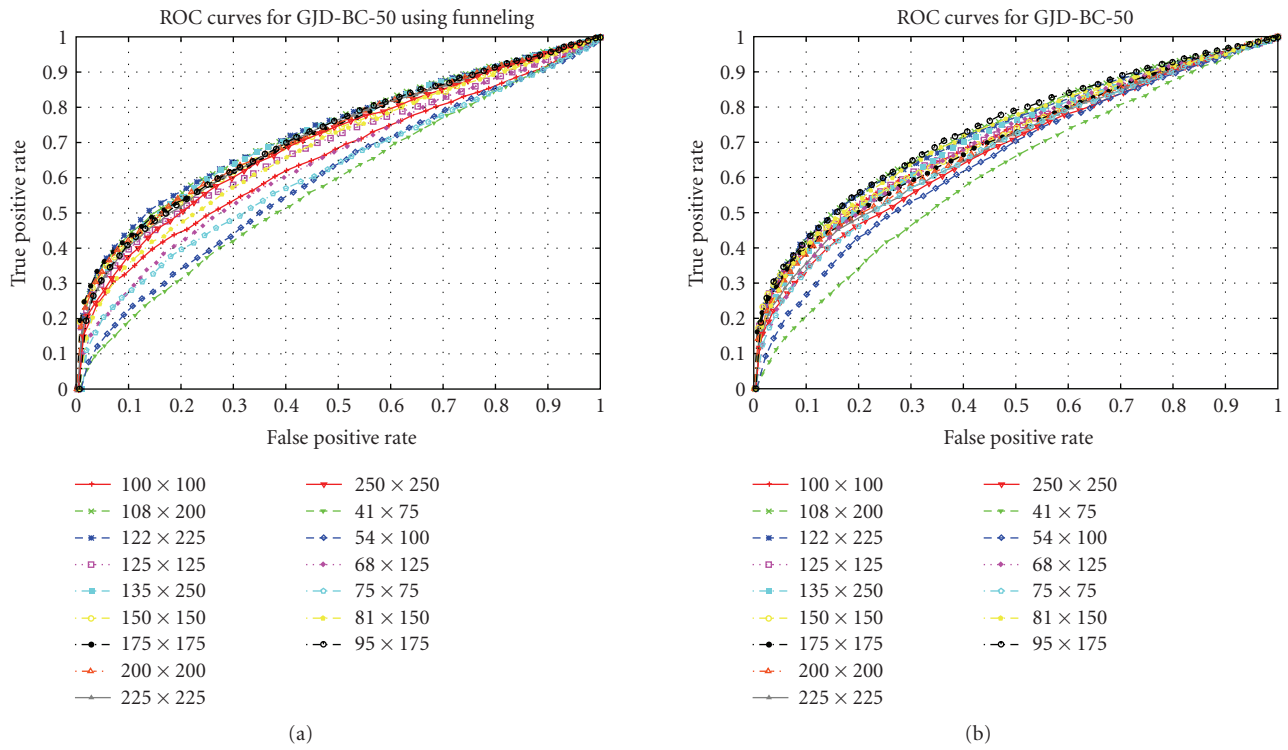


FIGURE 6: Effect of the image's region size on the performance of the GJD-BC-50 method. (a) Faces aligned using funneling; (b) unaligned faces.

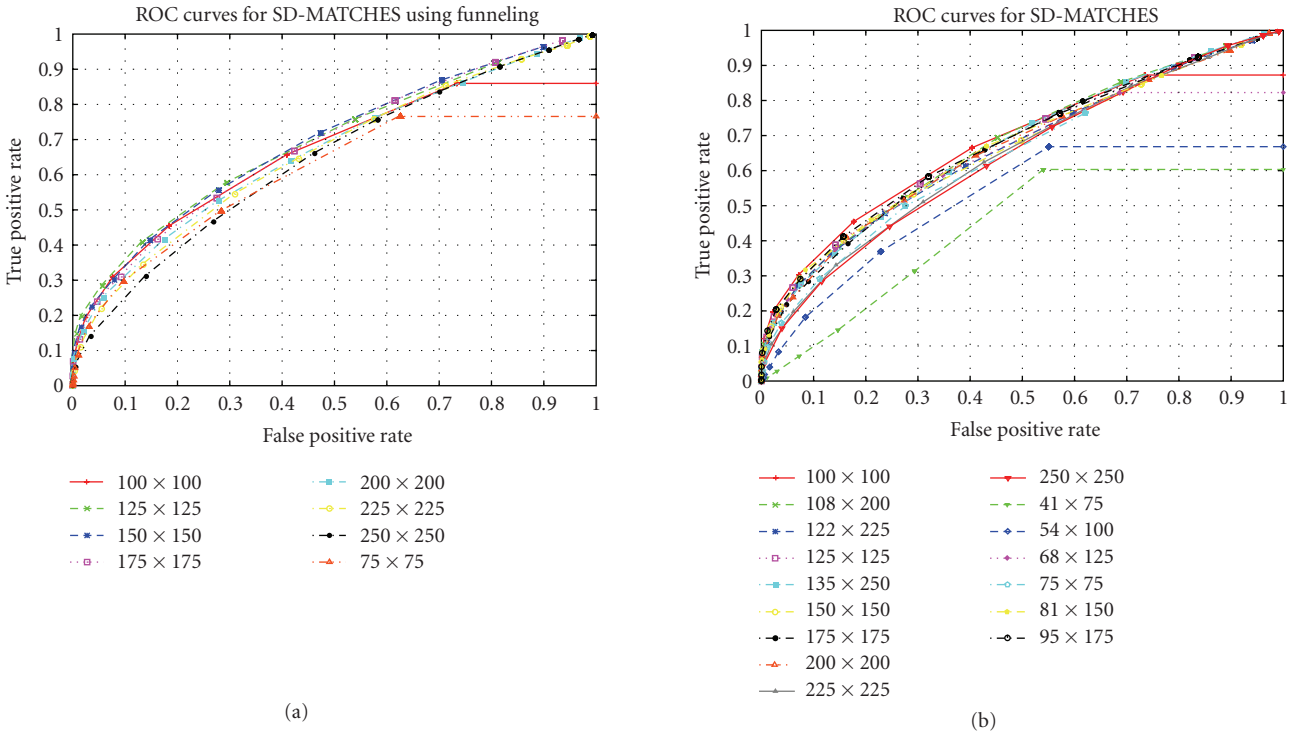


FIGURE 7: Effect of the image's region size on the performance of SD-MATCHES method. (a) Faces aligned using funneling, (b) unaligned faces.

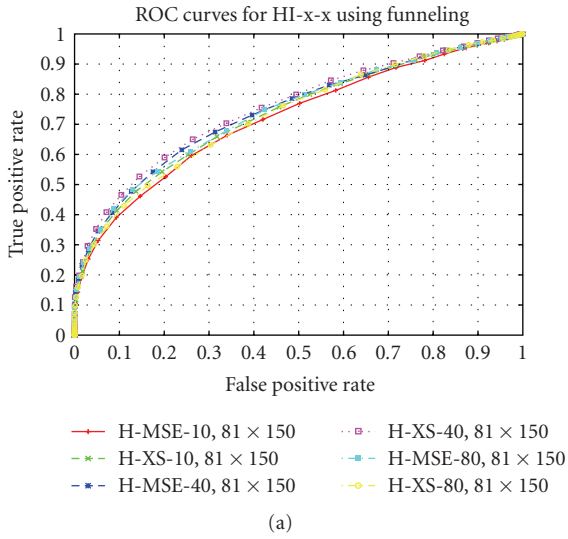


Figure 8: Comparison of the best working flavors of each method when funneling is used: (a) H-X-X, (b) GJD-BC, (c) SD-MATCHES.

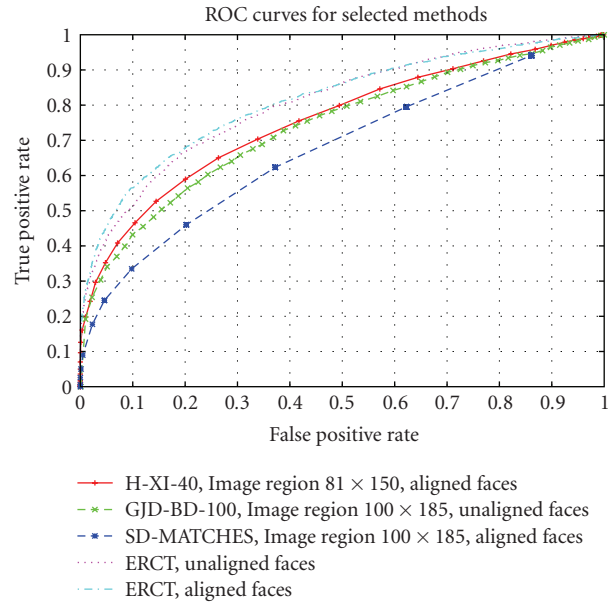


Figure 9: ROC curves of the best working variant of each method. Experiments were performed on faces aligned using funneling.

background information being analyzed on the performance of all algorithms. In Figure 4, the different amounts of face and background information that each image's size includes can be observed. The second thing is that in all cases (independently of the distance's measure and the method's parameters), small region sizes present the worst results, followed by the largest region sizes. Best results are obtained using medium-size regions.

Figure 5 shows the results for HI-XS-40. Best results are obtained for aligned images of size 81 × 150 (see also Figure 8(a)), which contains some background, but not very much (see Figure 4(c)). In the case of unaligned images, best results are obtained for images of size 95×175. Similar results were obtained when using 10 and 80 image divisions. For a fixed number of divisions, the Chi-Square measure works better than the mean square error (results not shown for space reasons).

Figure 6 shows the results for GJD-BC-50. Best results are obtained for aligned images of size 122 × 225 (see Figure 4(d)). In the case of unaligned images, best results are obtained for images of size 95 × 175. The most important thing that must be noticed here (see also Figure 8(b)) is that when the optimal image size is used and aligned faces are considered, using 10, 50, or 100 reference images; very similar results are given (in terms of MCA 0.6838, 0.6838, and 0.6847, resp.). This also holds when unaligned faces are used, but the difference is slightly larger (in terms of MCA 0.6752, 0.6780, and 0.6808, resp.). The experiments with reference sets of 10 and 100 images are not shown for space reasons.

Figures 7 and 8(c) show the results for SD-MATCHES. Best results are obtained again for aligned images; in this case a size of 125 × 125 gives better results (see Figure 4(e)). In the case of unaligned images, best results are obtained for

Table 3: Correct classification rates (LFW database, restricted setting). Experiments were performed on cropped regions of size 100 × 185 (**nw** = 2.4 and **nh** = 4.4), except for ERCF that considers the full image. MCA: *Mean classification accuracy.* SME: *Standard error of the mean.* In bold are the best results of each method.

	Without alignment		With alignment (funneling)	
Method	MCA	SME	MCA	SME
H-MSE-10	0.6375	0.0049	**0.6585**	0.0046
H-XS-10	0.6500	0.0043	**0.6668**	0.0044
H-MSE-40	0.6217	0.0055	**0.6527**	0.0057
H-XS-40	0.6383	0.0064	**0.6650**	0.0059
H-MSE-80	0.6527	0.0047	**0.6725**	0.0032
H-XS-80	0.6532	0.0053	**0.6785**	0.0055
GJD-EU	**0.6410**	0.0084	0.6375	0.0071
GJD-BC-10	**0.6777**	0.0080	0.6753	0.0082
GJD-BC-50	**0.6770**	0.0075	0.6742	0.0061
GJD-BC-100	**0.6798**	0.0065	0.6762	0.0069
SD-MATCHES	0.6015	0.0049	**0.6215**	0.0036
SD-SIMPLE	**0.6295**	0.0071	0.6288	0.0051
ERCF (from [15])	0.7245	0.0040	**0.7333**	0.0060

Table 4: Correct classification rates of the best methods (LFW database, restricted setting). MCA: *Mean Classification Accuracy.* SME: *Standard Error of the Mean.*

Method	Region Size	MCA	SME
SD-MATCHES, aligned faces	125 × 125	0.6410	0.0062
H-XS-40, aligned faces	81 × 150	0.6945	0.0048
GJD-BC-100, aligned faces	122 × 225	0.6847	0.0065
ERCF aligned faces (from [15])	250 × 250	0.7333	0.0060

Table 5: Processing Time. Time measures are in milliseconds. We carried out the experiments on a computer running Linux with an Intel Core 2 Duo E6750 2.66 GHz (2 GB RAM). FET/MT: Feature Extraction/Matching Time.

Method	H	H	H	GJD	GJD	GJD	GJD	SD	ERCF
Parameters	X-10	X-40	X-80	BC-1	BC-10	BC-50	BC-100	X	From [14]
FET (ms)	2.45	2.45	2.45	62	62	62	62	4.7	—
MT (ms)	0.033	0.118	0.230	0.37	2.63	5.59	15.55	64.7	2000
Image size	81 × 150			122 × 225				125 × 125	100 × 185

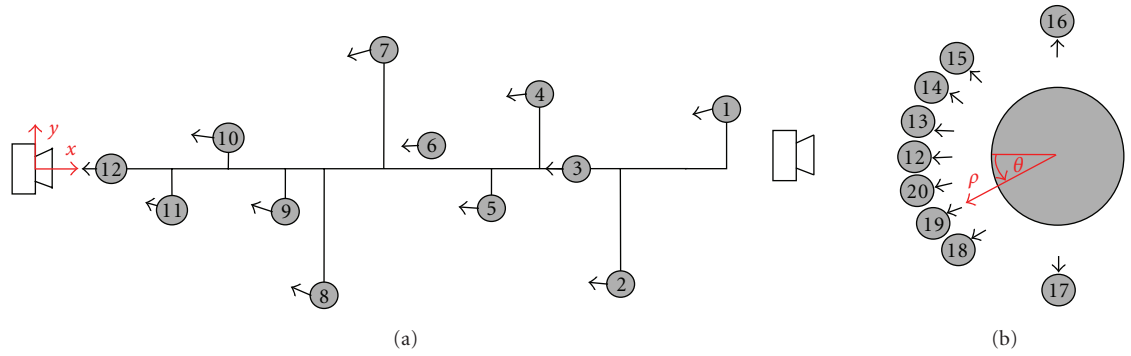


Figure 10: Experimental setup for image acquisition at different (a) distances and (b) angles. Arrows indicate the angular pose of the subjects. (a) Cartesian coordinates of acquisition points (in centimetres) relative to the camera's focus: P1 (1088,90), P2 (906,−180), P3 (785,0), P4 (755,151), P5 (665,−51), P6 (574,30), P7 (514,181), P8 (423,−181), P9 (332,−61), P10 (272,30), P11 (181,−61), and P12 (90,0). (b) Polar coordinates of acquisition points (radius in centimetres and angles in degrees) relative to the camera's focus: P16 (90,90°), P15 (90,45°), P14 (90,30°), P13 (90,15°), P20 (90,−15°), P19 (90,−30°), P18 (90,−45°), and P17 (90,−90°).

images of size 100×100. In all cases, best results are obtained with the SD-Matches variant. Very low performance results are obtained by the SD-Simple variant.

Finally, Figure 9 shows the ROC curves of the best variant of each method.

4.4. Discussion. If one analyzes the performance obtained by the different methods, ERCF obtains clearly the best results (see Table 4). Best LBP-based method (H-XS-40) is almost 3.9% below ERCF, and about 1% over GJD's best method (GJD-BC-100). However, if one now analyzes the processing speed of the methods, the best variant of LBP-based methods (H-XS-40) is at least 400 times faster than ERCF (see Table 5), and 30 times faster than the best Gabor method (GJD-BC-100). The high processing time of ERCF and GJD can be too restrictive for some applications, in particular in the ones that require real-time operation (e.g., HRI) as well in applications where very large amounts of data are being analyzed (e.g., search in a very large multimedia database). The Borda count ranking of each of the features is the slowest operation of GJD, while the slowest part of ERCF corresponds to the computation of the normalized cross-correlation in the selection of pairs of regions to be quantized using ERCF. However, it should be noted that in a face identification scenario, as the one reported for the FERET case, it is not required that GJD use a reference set. In this case (GJD-BC-1 in Table 5), the method needs about 63 milliseconds to analyze a face image.

It is also interesting to analyze which kind of information uses each method by looking at the optimal regions they use. The regions are shown in Figure 4 as well as the normalized width and height of the face images (last row). SD methods, specifically SD-Matches that has an optimal region size of 125×125 (see Figure 4(e)), show better performance when there is the fewest possible background in the image, but without removing any part of the face. The methods need as much as possible face information to obtain a correct matching. However, the background disturbs the matching process (a face keypoint could be matched to a background keypoint). In the case of LBP-based methods, specifically for H-XS-40 that has an optimal region size of 81×150 (see Figure 4(c)), it seems that some background but not much helps. Probably, this additional information about the face's contour helps the recognition process. Finally, in the case of the GJD methods, specifically GJD-BC-100 that has an optimal region size of 122×225 (see Figure 4(d)), the image contains much more background. The reason is twofolds: (i) the Gabor-filters encode information about the contour of the face, and (ii) large regions allow the use of large filters, which encoded large-scale information.

It is important to compare the optimal region sizes of the methods in the LWF, with the sizes used in the FERET experiments. However, it should be noted that the images in both databases have different resolutions. Therefore, instead of comparing region sizes, normalized image's width (nw) and height (nh) need to be used. In our FERET experiments, the nw/nh values are 1.6/3.0 for the 100×185 case, and 3.0/3.7 for the 203×251 case. The nw and nh values of the optimal region sizes, in the LWF case, are shown in Figure 4. By comparing these normalized values we observe that there is a concordance. (i) SD and GJD methods behave much better when normalized sizes of 3.0/3.7 are used in FERET, and in LFW behave better with values of 3.0/3.0 for the case of the SD method, and 2.9/5.4 for the case of the GJD method. (ii) In the case of the H-XS-40 method, similar results are obtained in FERET with 1.6/3.0 or 3.0/3.7, which is concordant with the selected values of 2.4/4.4 in LWF. Naturally, the normalized values in both databases are not the same, because in the FERET case we decided to use just two, fixed image's sizes, while in the LWF we allow the methods to choose the best values.

Finally, it is interesting to analyze how much the methods' performance depends on the alignment's accuracy. By observing Table 3, it can be seen that the methods with the largest dependence on the alignment's accuracy are the H-X-X. These results are consistent with the one obtained in the FERET database. On the opposite site, SD methods are very robust to alignment errors, which is also consistent with the results obtained in the FERET case. As it can be noticed, GJD performs worse when alignment is used. We think this is related with the way in which the used alignment method (funneling) works. Funneling aligns the whole face (shape), and not the eyes. As observed in the results obtained for FERET, GJD seems to be very sensible to good eyes' alignment.

5. Comparative Study Using Real HRI Database

The UCHFaceHRI database was built with the goal of allowing the study of face analysis methods in tasks such as detection, recognition, and relative pose determination of humans using face information, for HRI (Human-Robot Interaction) applications. The database contains images from 30 individuals, which were taken in 20 different relative camera-individual poses (see acquisition points in Figure 10), in outdoor and in indoor settings, at a resolution of 1024×768 pixels. Five different face expressions were considered for the case of the frontal face (P12 acquisition point): neutral expression, surprised, angry, sad, and happy. Thus, the database contains 48 images for each individual. Each of these 48 face images is specified as Fjkl, where j indicates that the image was taken at the acquisition point Pj and k indicates which expression is associated to this image (neutral: $k = a$, surprised: $k = b$, angry: $k = c$, sad: $k = d$, happy: $k = e$). This index is valid only in the case of images taken in the acquisition point P12. Finally, l indicates if the image was taken in an indoor ($l = i$) or an outdoor ($l = o$) environment. Figure 11 shows the 24 indoor images corresponding to a given individual. The database can be downloaded in [30].

In all experiments the F12ai face images composed the gallery set. We define 14 specific and global test sets, to analyze the methods' invariance to the scale, orientation, and expression of the faces, considering indoor and outdoor illumination conditions as follows.

(i) Scale test sets. S-I: Scale Indoor (images F10i-F11i), S-O: Scale Outdoor (images F10o-F12o).

(ii) Expression test sets. E-I: Expression-Indoor (images F12bi-F12ei), E-O: Expression-Indoor (images F12bo-F12eo).

(iii) Rotation test sets. R-I: Rotation-Indoor (images F13i, F14i, F19i, F20i), R-I/15: Rotation-Indoor in 15 degrees (images F13i, F20i), R-I/30: Rotation-Indoor in 30 degrees (images F14i, F19i), R-O: Rotation-Outdoor (images F13o, F14o, F19o, F20o), R-O/15: Rotation-Outdoor in 15 degrees (images F13o, F20o), R-O/30: Rotation-Outdoor in 30 degrees (images F14o, F19o).

(iv) Global test sets. Scale: S = S-I + S-O, Expression: E = E-I + E-O, Rotation: R = R-I + R-O, Global Indoor: G-I = S-I + E-I + R-I, Global Outdoor: G-O = S-O + E-O + R-O, and Global: G = D + E + R = G-I + G-O.

In the experiments we considered the best working variants (distance's measure and region's size) of each method (H, GJD-BC, SD, and ERCF), according to the results obtained in LFW. To have the same conditions than in the LWF experiments, the faces were aligned using the annotated eyes, and the cropping was done without using funnelling, but using the estimated bounding box that would have been obtained if funnelling was used. This estimation was obtained by measuring the eyes positions of a subset of 20 LWF-funnelled images. As in the case of LWF, the distance between eyes was 42 pixels.

For the evaluation of ERCF, we trained a system using the implementation of the author (available on http://lear.inrialpes.fr/people/nowak/similarity/index.html) of ERCF and the same parameters used to obtain the results presented in [15], which were obtained by a direct communication with the authors of the LFW database. For ERCF we are presenting results for four cases, each one corresponding to a different value of C when training the SVM classifier. The results presented in the previous section for ERCF correspond to $C = 1$. Here we used as training set, the complete test set of LFW (6000 pairs of images).

Table 6 shows the top-1 recognition rates obtained in these tests. Main conclusions are as follows.

(i) Comments on indoor/outdoor tests are as follows.

(a) For all methods, much better results are obtained for indoor faces than for outdoor faces. This is a clear indication that the analyzed methods are not robust to outdoor illumination. Some improvement may be achieved if preprocessing stages are added.

(b) H-X-X methods obtain the highest recognition rate with outdoor faces, followed by GJD and ERCF.

(c) SD performance is strongly affected by outdoor illumination.

(ii) Comments about Scale tests are as follows.

(a) The best performing method is H-X-X, followed by GJD, ERCF, and SD, in that order. However, if we consider only indoor images (S-I set), the best performing methods are H-XS-40 and ERCF, followed by the SD-variants. GJD got the lowest top-1 RR.

(b) In the case of outdoor images all methods have a very low performance, with the best ones (H-HI-40 and H-MSE-40) achieving only a 50% top-1 RR.

(iii) Comments about Expression tests:

(a) HI-X-X shows the best performance followed by ERCF. In the third place comes GJD followed by SD. The same holds if we consider only indoor images (E-I).

(b) In the case of outdoor images, all methods have a very low performance, with the best one (H-XS-40) achieving only a 50.7% top-1 RR.

(iv) Comments about rotation tests are as follows.

(a) Which methods is the best depends on the amount of rotation in the images and on the illuminations conditions. In case of low rotations (15 degrees) with indoor or outdoor illumination, HI-X-X got the highest top-1 RR. In case of higher rotations (30 degrees) and indoor illumination, the same happens. However, in case of 30 degrees rotation and outdoor illumination, ERCF got the top-1 RR.

(b) In indoor conditions, SD is more robust to rotations than GJD. Moreover, SD-Matches and SD-Simple present the second best results in some indoor image cases. However, in outdoor conditions their performance is quite low.

(c) In general terms, the performance of some methods in indoor images with 15 degrees rotation is acceptable (~76%). However, no method gives acceptable results for outdoor images with low rotation (15 degrees), or for rotations in 30 degrees.

(v) Comments about global results are as follows.

(a) Overall, best results are obtained in most of the cases by one of the HI-X-X variants (7 out of 8 subset test, S-I, S-O, E-I, E-O, R-I/15, R-I/30, R-O/15). The second best method is ERCR (being the best in R-O/30 and the second best in most of the cases). If we consider only indoor conditions, GJD and SD got a similar performance, with one of the SD variants (SD-Simple) obtaining slightly better results than GJD. However, if both indoor and outdoor images are considered, the third best method is GJD.

Table 6: UCHFaceHRI tests. Top-1 recognition rate. Experiments are performed with detected eyes. In bold are the best results for each condition. Methods that have differences of 1% or less are considered as having the same performance. See main text for a description about the different experiments.

Method	S-I	S-O	S	E-I	E-O	E	R-I/15	R-I/30	RI	R-O/15	R-O/30	RO	R	G-I	G-O	G
H-HI-40	95.0	**50.0**	**68.0**	**92.5**	48.7	**68.1**	**75.9**	**32.8**	**54.3**	**53.4**	24.1	**38.8**	**46.6**	**78.0**	**45.8**	**60.4**
H-MSE-40	95.0	**50.0**	**68.0**	**92.5**	48.7	**68.1**	**75.9**	**32.8**	**54.3**	**53.4**	24.1	**38.8**	**46.6**	**78.0**	**45.8**	**60.4**
H-XS-40	**98.3**	45.6	66.7	89.2	**50.7**	**67.8**	69.0	27.6	48.3	51.7	24.1	**37.9**	43.1	75.0	**45.2**	58.7
GJD-BC-F	85.0	38.9	57.3	73.3	48.0	59.3	43.1	10.3	26.7	36.2	15.5	25.9	26.3	57.4	38.5	47.1
SD-SIMPLE	91.7	11.1	43.3	61.7	9.3	32.6	56.9	15.5	36.2	5.2	3.4	4.3	20.3	57.8	8.1	30.7
SD-FULL	86.7	6.7	38.7	63.3	8.0	32.6	34.5	6.9	20.7	5.2	5.2	5.2	12.9	51.4	6.7	27.0
SD-MATCHES	88.3	12.2	42.7	51.7	8.0	27.4	46.6	27.6	37.1	10.3	6.9	8.6	22.8	53.4	9.3	29.3
ERCF C = 1e-06	96.7	35.6	60.0	76.7	40.0	56.3	36.2	29.3	32.8	46.6	**27.6**	37.1	34.9	63.5	37.9	49.5
ERCF C = 0.0001	96.7	42.2	64.0	80.0	42.0	58.9	46.6	31.0	38.8	43.1	**27.6**	35.3	37.1	67.2	39.9	52.3
ERCF C = 0.1	**98.3**	24.4	54.0	74.2	30.0	49.6	56.9	20.7	38.8	34.5	15.5	25.0	31.9	65.2	27.0	44.3
ERCF C = 1	**98.3**	24.4	54.0	74.2	30.0	49.6	56.9	20.7	38.8	34.5	15.5	25.0	31.9	65.2	27.0	44.3

6. Comparative Study Using FRGC

From the reported experiments it can be observed that the methods that perform better in our experiments are the LBP-based (H-X-X) and Gabor-based (GJD) ones. These methods are further analyzed using the FRGC ver2.0 database [17]. This database consists of 50,000 face images divided into training and validation partitions. In our experiments the training partition was not used, because one of our main requirements is that methods under comparison should be fully on-line. The validation partition consists of data from 4,003 subject sessions. A subject session consists of controlled and uncontrolled images. The controlled images were taken in a studio setting, and they are full frontal facial images taken under two lighting conditions and with two facial expressions (smiling and neutral), while the uncontrolled images were taken in varying illumination conditions [17]. Each set of uncontrolled images contains two expressions, smiling and neutral. In our analysis we will focus on two FRGC tests: *Experiment 1*, which corresponds to a control experiment where the gallery and the probe sets consist of controlled still images, and *Experiment 4*, which measures recognition performance from uncontrolled images (the probe set consist of single uncontrolled still images; the gallery is composed by controlled still images).

Figure 12 shows the ROC curve obtained in experiment 1 by the best methods under comparison. It should be stressed that in our test we have used all possible image pair comparisons that can be carried out in experiment 1 ($16,028 \times 16,028$), and not the image pairs defined by the ROC I–ROC III FRGC subexperiments that some papers report. As it can be observed the obtained results are concordant with the ones of similar reported approaches , for instance in [31, 32]. But, if we compare these methods with recent kernel-based approaches, as the ones proposed by Liu [33] (Gabor-Multiclass-KFDA) or Zhao et al. (LBP KFDA) [31], we observe that kernel approaches obtain much higher results than LBP- or Gabor-based approaches, about 10% higher verification rate for a given FAR. However, it should be remembered that the kernel approaches need to be trained in the database, and they are much slower than the methods under comparison. From Figure 12 it is also interesting to note the dependency of the LBP-based methods' performance on the number of partitions. Methods using a larger number of partitions get better results than methods using a smaller number of partitions. This phenomenon although being logic was not clearly observed in the other databases. Probably with very large database the number of partitions is an important parameter to be considered.

We also analyzed the methods under comparison using the FRGC, experiment 4. By analyzing the results, similar conclusions were obtained: (i) the results are concordant with the ones of similar approaches reported in the literature (see, e.g., [26]), (ii) kernel approaches get much better results, and (iii) the performance of the LBP-based methods depends on the number of partitions.

7. Discussion and Conclusions

In this article, a comparative study among face-recognition methods in unconstrained environments was presented. The analyzed methods were selected by considering their suitability for the defined requirements—real-time operation, just one image per person, fully on-line (no training), robust behavior in unconstrained environments, and their performance in former studies. The comparative study was carried out using three databases: FERET, LFW, and UCHFaceHRI. The well-known FERET database was used as a baseline for comparison, and experiments were carried out in different subsets that include variations in illumination, nonaccurate eye's annotations, and occlusions. The LFW database implicitly includes aspects such as scale, pose, lighting, focus, resolution, facial expression, accessories, makeup, occlusions, background, and photographic quality, while the UCHFaceHRI explicitly includes aspects such as scale (distance to the camera), expressions (neutral, surprised, angry, sad, and happy), pose (0, ± 15 , and ± 30 degrees of

Figure 11: UCHFaceHRI database. Examples of 24 indoor images corresponding to an individual. The face-image F*jki* corresponds to an image taken acquisition point *j* (see Figure 10). *i* stands for indoor. In the case of the F12*k*i images, the *k* index means: (a) Neutral expression, (b) Surprised, (c) Angry, (d) Sad, and (e) Happy.

out-of-plane rotation), and illumination (indoor/outdoor). The methods under comparison are generalized-PCA, LBP histograms, Gabor Jets descriptors, SIFT descriptors, and ERCF. We will comment about the main results of this study, and we will draw some conclusions of this work.

Comments on the Size of the Face Region. What was very surprising to us is the large dependence of the methods to the amount of face and background information that is included in the face's images. This effect was clearly seen in our FERET and LFW experiments. For instance, in the FERET case, SD increases its recognition rate in more than 20% depending on the size of the face images. In the LWF case where experimental conditions are much harder, LBP-based methods and SD increase their recognition rates in ~4%, depending on the size of the face images. We also observe that the different methods have different requirements. LBP-based methods concentrate themselves mostly in the face area, but it seems that additional information about the face's chin, which is only observed if some background is included in the images, helps the recognition process. On the other hand, GJD methods need much more background. The reason is twofolds: (i) the Gabor filters encode information about the contour of the face, and (ii) large regions allow the use of large filters, which encoded large-scale information. SD methods show better performance when there is the fewest possible background in the image, but without removing any part of the face. The methods needs as much as possible face information to obtain a correct matching, but the background disturbs this process (a face keypoint could be matched to a background keypoint).

Comments on the Illumination Conditions. Most of the methods behave very well in natural, indoor illumination conditions, the exception being SD. This can be clearly seen in the FERET experiments (*fa-fc*). However, this situation changes drastically with outdoor illumination conditions. The performance of all methods decreases largely with outdoor illumination. Clearly, face recognition in outdoor conditions is still a nonsolved problem.

Comments on Pose Variations. Invariance against pose variations is a second main problem in face recognition. In

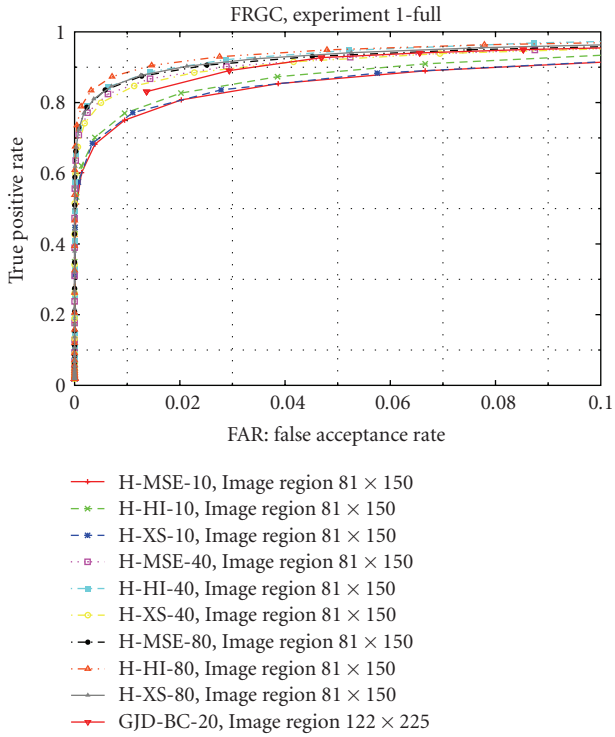


Figure 12: ROC curves of the best methods under comparison in FRGC, experiment 1.

the UCHFaceHRI experiments it can be observed that yaw rotations in 15 degrees affect largely the performance of all methods; the recognition rates decrease in more than 20%. In the 30-degrees case the situation is even worse, the recognition rates fall in more than 60%. In relation, we also believe that the main reason for the low results that are obtained in the LFW database is due to the variations in the faces' pose.

Comments on Alignment, Occlusions, and Expressions. From our experiments we conclude that the analyzed methods are robust to inaccurate alignment, face occlusions, and variations in expressions, to a large degree. Accepting that these factors affect the face-recognition process, their influence in the algorithms' performance is much lower than outdoor illumination or pose's variations.

Conclusions about the Performance of Methods. The question of which method is the best is a very difficult one. However, we could say that LBP-based methods are an excellent election if we need real-time operation as well as high recognition rates. In the UCHFaceHRI experiments some of the LBP variants got the best results, while in the LWF case they got the second best results.

Gabor-based methods are also an adequate election. Although they got a lower performance in UCHFaceHRI than LBP-based methods, they got a similar performance in LFW, and slightly better results in FERET. However, Gabor-based methods are slower than LBP ones. Probably some work can be done to develop strategies that select which filters to use (some research in this direction has been reported in [10]). A last interesting aspect to be mentioned is that the proposed strategy of using a reference set of images in the case of comparing pairs of images was successful and better than using the Euclidian distance.

ERCF is a novel and promising matching method. However, it has some drawbacks, the first one being its low processing speed, which does not allow its application in real-time conditions. Moreover, the method has several parameters, and it seems that its performance depends on the correct selection of them. Thus, although the method achieves the best results in the LFW database, being clearly superior to the others, it got the second place in the UCHFaceHRI experiments. In these experiments LBP-based methods work better than ERCF, in particular in difficult cases such as outdoor images, out-of-plane rotation, and facial expressions. This may be due to the fact that the learning done by ERCF does not generalize as the results reported for LFW seem to indicate. This may be due to the fact that the images from LFW were obtained from news images, which in general are taken by professional photographers, and therefore are obtained under good illumination, and because they are also taken in indoor conditions, which are the cases where ERCF works best.

SD methods performed very well in some of our experiments, achieving similar recognition rates than LBP-based and Gabor-based methods. However, SD methods have a large dependence to illuminations conditions. This is especially true for the case of outdoor illumination, were the methods' performance decrease largely. It is interesting to note that the large dependence of SD methods to illumination conditions is not clearly reported in the SIFT-related literature.

The generalized PCA method got the worse results in the FERET experiments and was not further analyzed in this study. We believe that under the main requirements of this study (real-time operation, just one image per person, and no training stages), eigenspace-based holistic methods are not competitive against the other methods.

When the best methods under analysis are compared against novel kernel-based approaches [31, 33] (e.g., in the FRGC database), they obtain a lower performance. However, it should be noted that kernel-based methods are intended to be used in other kinds of applications, which do not have the requirements of real-time and full on-line operation.

Future Work. We believe that still there are many aspects that can be improved in the recognition of faces in unconstrained environments. However, in the medium term, we will concentrate on: (i) the analysis of pre-processing algorithms and other strategies to achieve invariance against outdoor illumination conditions, (ii) the combined use of methods (e.g., ERCF and LBP-based or kernel-based and LBP) that can allow achieving, at the same time, high recognition rates and processing speed, (iii) the study of the influence of face's resolution in the recognition process, and (iv) a more deep analysis of the facial expression effect in the recognition of faces.

References

[1] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: a literature survey," *ACM Computing Surveys*, vol. 35, no. 4, pp. 399–458, 2003.

[2] X. Tan, S. Chen, Z.-H. Zhou, and F. Zhang, "Face recognition from a single image per person: a survey," *Pattern Recognition*, vol. 39, no. 9, pp. 1725–1745, 2006.

[3] R. Chellappa, C. L. Wilson, and S. Sirohey, "Human and machine recognition of faces: a survey," *Proceedings of the IEEE*, vol. 83, no. 5, pp. 705–740, 1995.

[4] A. F. Abate, M. Nappi, D. Riccio, and G. Sabatino, "2D and 3D face recognition: a survey," *Pattern Recognition Letters*, vol. 28, no. 14, pp. 1885–1906, 2007.

[5] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[6] A. S. Mian, M. Bennamoun, and R. Owens, "An efficient multimodal 2D-3D hybrid approach to automatic face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 11, pp. 1927–1943, 2007.

[7] R. Singh, M. Vatsa, and A. Noore, "Integrated multilevel image fusion and match score fusion of visible and infrared face images for robust face recognition," *Pattern Recognition*, vol. 41, no. 3, pp. 880–893, 2008.

[8] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: a database for studying face recognition in unconstrained environments," Tech. Rep. 07-49, University of Massachusetts, Amherst, Mass, USA, October 2007.

[9] J. Ruiz-del-Solar and P. Navarrete, "Eigenspace-based face recognition: a comparative study of different approaches," *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, vol. 35, no. 3, pp. 315–325, 2005.

[10] J. Zou, Q. Ji, and G. Nagy, "A comparative study of local matching approach for face recognition," *IEEE Transactions on Image Processing*, vol. 16, no. 10, pp. 2617–2628, 2007.

[11] J. Ruiz-del-Solar and J. Quinteros, "Illumination compensation and normalization in eigenspace-based face recognition: a comparative study of different pre-processing approaches," *Pattern Recognition Letters*, vol. 29, no. 14, pp. 1966–1979, 2008.

[12] M. Correa, J. Ruiz-del-Solar, and F. Bernuy, "Face recognition for human-robot interaction applications: a comparative study," in *Proceedings of the RoboCup International Symposium*, Lecture Notes in Computer Science, Suzhou, China, July 2008.

[13] T. Ahonen, A. Hadid, and M. Pietikäinen, "Face description with local binary patterns: application to face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 12, pp. 2037–2041, 2006.

[14] F. Moosmann, E. Nowak, and F. Jurie, "Randomized clustering forests for image classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 9, pp. 1632–1646, 2008.

[15] Labeled Faces in the Wild database, "Results," http://vis-www.cs.umass.edu/lfw/results.html.

[16] P. J. Phillips, P. J. Flynn, T. Scruggs, et al., "Overview of the face recognition grand challenge," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '05)*, vol. 1, pp. 947–954, San Diego, Calif, USA, June 2005.

[17] Face Recognition Grand Challenge, http://www.frvt.org/FRGC.

[18] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '01)*, vol. 1, pp. 511–518, Kauai, Hawaii, USA, December 2001.

[19] R. Verschae, J. Ruiz-del-Solar, and M. Correa, "Face recognition in unconstrained environments: a comparative study," in *Proceedings of the Workshop on Faces in Real-Life Images: Detection, Alignment, and Recognition (ECCV '08)*, pp. 1–12, Marseille, France, October 2008, CD Proceedings.

[20] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[21] J. Ruiz-del-Solar, P. Loncomilla, and Ch. Devia, "A new approach for fingerprint verification based on wide baseline matching using local interest points and descriptors," in *Proceedings of the 2nd IEEE Pacific Rim Symposium on Image and Video Tecnology (PSIVT '07)*, vol. 4872 of *Lecture Notes in Computer Science*, pp. 586–599, Santiago, Chile, December 2007.

[22] J. Ruiz-Del-Solar, Ch. Devia, P. Loncomilla, and F. Concha, "Offline signature verification using local interest points and descriptors," in *Proceedings of the 13th Iberoamerican Congress on Pattern Recognition (CIARP '08)*, vol. 5197 of *Lecture Notes in Computer Science*, pp. 22–29, Havana, Cuba, September 2008.

[23] M. Bicego, A. Lagorio, E. Grosso, and M. Tistarelli, "On the use of SIFT features for face authentication," in *Proceedings of the Conference on Computer Vision and Pattern Recognition Workshop (CVPRW '06)*, p. 35, New York, NY, USA, June 2006.

[24] B. Fröba and A. Ernst, "Face detection with the modified census transform," in *Proceedings of the 6th IEEE International Conference on Automatic Face and Gesture Recognition (FGR '04)*, pp. 91–96, Seoul, Korea, May 2004.

[25] P. J. Phillips, H. Wechsler, J. Huang, and P. J. Rauss, "The FERET database and evaluation procedure for face-recognition algorithms," *Image and Vision Computing*, vol. 16, no. 5, pp. 295–306, 1998.

[26] X. Tan and B. Triggs, "Fusing Gabor and LBP feature sets for kernel-based face recognition," in *Proceedings of the 3rd International Workshop on Analysis and Modeling of Faces and Gestures (AMFG '07)*, vol. 4778 of *Lecture Notes in Computer Science*, pp. 235–249, Rio de Janeiro, Brazil, October 2007.

[27] Y. Su, S. Shan, X. Chen, and W. Gao, "Patch-based Gabor fisher classifier for face recognition," in *Proceedings of the 18th International Conference on Pattern Recognition (ICPR '06)*, vol. 2, pp. 528–531, Hong Kong, August 2006.

[28] S. Shan, W. Zhang, Y. Su, X. Chen, and W. Gao, "Ensemble of piecewise FDA based on spatial histograms of local (Gabor) binary patterns for face recognition," in *Proceedings of the 18th International Conference on Pattern Recognition (ICPR '06)*, vol. 4, pp. 606–609, Hong Kong, August 2006.

[29] G. B. Huang, V. Jain, and E. Learned-Miller, "Unsupervised joint alignment of complex images," in *Proceedings of the 11th IEEE International Conference on Computer Vision (ICCV '07)*, pp. 1–8, Rio de Janeiro, Brazil, October 2007.

[30] UCHFaceHRI database, http://vision.die.uchile.cl/2/Databases.htm.

[31] J. Zhao, H. Wang, H. Ren, and S. C. Kee, "LBP discriminant analysis for face verification," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '05)*, vol. 3, pp. 167–172, San Diego, Calif, USA, June 2005.

[32] H. Yang and Y. Wang, "A LBP-based face recognition method with hamming distance constraint," in *Proceedings of the 4th*

International Conference on Image and Graphics (ICIG '07), pp. 645–649, Chengdu, China, August 2007.

[33] C. Liu, "Capitalize on dimensionality increasing techniques for improving face recognition grand challenge performance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 5, pp. 725–737, 2006.