Face recognition using multiple interest point detectors and SIFT descriptors

C. Fernandez and M.A. Vicente
Miguel Hernandez University
Av. Universidad s/n. 03202 Elche (Spain)
{c.fernandez,suni}@umh.es

Abstract

The use of interest point detectors and SIFT descriptors for face recognition is studied in this paper. There are two main novelties with respect to previous approaches using SIFT features. First, the use of two scale-invariant interest point detectors (namely, Harris-Laplace and Difference of Gaussians) which are combined in order to detect both corner-like structures and blob-like structures in face images. Second, the distance measure used, which takes into account both the number of matching points found between two images (according to their SIFT descriptors) and the coherence of these matches in terms of scales, orientations and spacial configuration. The results obtained with our model-based algorithm are compared with those of a classic appearance-based face recognition method (PCA) over two different face databases: the well-known AT&T database and a face database created at our University.

1. Introduction

Face recognition has been an area of active research over the last 40 years [27]. The most popular face recognition approaches make use of appearance-based projection methods, like PCA (principal component analysis) [25] [12], ICA (independent component analysis) [2] [8] [14] or LDA (linear discriminant analysis) [4]; while model-based methods have seldom been used [18] [7] [13] [11]. However, recent research on interest point detectors and invariant descriptors (mainly for object recognition applications) has open a new alternative for model-based face recognition and authentication. In this paper, we propose an algorithm based on Lowe's [17] SIFT descriptors (Scale Invariant Feature Transform descriptors).

SIFT descriptors were initially developed for object recognition purposes [15] and since then, they have been widely used for such purpose and for many others, like robot navigation [23], scene classification [21], etc. Lately, SIFT features have also been used by different authors in the field of face recognition, with promising results. However, recent publications in this field show that -so far- there is not an agreement on the best way to use such descriptors for face recognition, mainly when it comes to measure distances between face images from their SIFT features.

Having a look at some of the proposals found in the literature, in [16], Lowe proposes to use his SIFT features for face recognition in a similar way as they are used for object recognition. However, no details are given about how to compare the similarity of a certain test image to training images belonging to two different persons; and the problem of face authentication is not addressed either. In [20], although the purpose is not exactly face recognition, the authors measure the distance between two face images as the average value of the distances between all matches that fulfil certain geometrical constraints. In our opinion (and according to Lowe's papers) the absolute value of distances between matches is not as reliable as the ratio of the distances between the first and the second best matches, so SIFT features are not fully exploited in such proposal. In [24], where the goal is video retrieval, a combined approach is used, where PCA is used to locate eyes, nose and mouth; and SIFT features are used to describe 5 predefined face areas around such points. Tracks in a video sequence allow to construct sets of faces for the same person, which are represented by K-means based vector quantization and compared using χ^2 statistic. In [6], different options are compared: the first option is to measure the distance between two faces as the distance of the best matching pair of descriptors; while further options make use of previous knowledge about the location of eyes and mouth in the images in order to measure distances in predefined face areas.

Based on the above mentioned previous results from other research groups, one of the goals of this paper is to propose a new distance measure, which fully exploits the potential of SIFT features in the face recognition problem.

2. Interest point detectors

Interest point detection is a key factor for describing images from local features. Given a certain image, some relevant points have to be selected, so that a certain feature com-

978-1-4244-2154-1/08/$25.00 ©2008 IEEE

puted in the surroundings of such points can represent the image contents and therefore distinguish such image among many others.

A common requirement for all applications is repeatability: a good detector should find the same interest points in different images of the same object or person; where such images can differ in scale, orientation, viewpoint, lighting conditions, amount of noise, etc. Another requirement is descriptive power: the detected points should correspond to meaningful regions of the image. In this sense, the ideal detector depends on the application.

Focusing on the face recognition problem, an interest point detector should fulfil lighting and noise level invariance, scale invariance (as the distance from the camera may vary substantially from one image to another), orientation invariance (as the person may tilt the head in a plane perpendicular to the camera axis) and viewpoint invariance (as the user may not always be perfectly facing the camera). However, both orientation and viewpoint change are not expected to be high in this kind of applications.

Viewpoint invariance is usually addressed using affine invariant detectors, like those of Alvarez and Morales [1], Baumberg [3] or Mikolajczyk and Schmid [19]. However, this is only an approximation, as only planar surfaces suffer affine transformations under viewpoint changes. Human faces are not planar and therefore viewpoint invariance can not be obtained this way. Besides, simpler detectors are always faster and usually perform better than affine-invariant ones when the range of viewpoint change is small. As this is the case in face recognition applications, simpler detectors will be used in our algorithm.

Among these simpler detectors, there are two approaches which look for different features: the Harris-Laplace detector [19] and the SIFT or Difference of Gaussian detector [17]. The first one is a scale-invariant version of the well known Harris corner detector [9], and looks for corner-like or junction-like features in images. The second one is an approximation to the Laplacian of Gaussian operator and, due to the symmetry of this operator, looks mainly for blob-like features in the images. Both detectors will be briefly explained in the next sections.

2.1. Harris-Laplace detector

Harris-Laplace detector looks for points in the image whose value of cornerness is locally maximal. Cornerness is defined from the autocorrelation function, as eqs. 1 and 2 show. Areas of high cornerness correspond to corners or junctions in the images.

$$c(x, y; \Delta x, \Delta y) = [\Delta x \Delta y]\, Q(x, y) \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix} \qquad (1)$$

$$H = \lambda_1 \lambda_2 - 0.04(\lambda_1 + \lambda_2)^2 \qquad (2)$$

In equations 1 and 2, c is the autocorrelation function evaluated in a point (x, y) of the image, and the cornerness H is obtained from the eigenvectors λ_1 and λ_2 of $Q(x, y)$. In order to obtain scale invariance, a scale-space representation of the image is obtained by convolving it with Gaussian kernels of different size, and local maxima of cornerness are detected at each scale. For each of these local maxima, a search is performed over all scales to find the maxima over scales of the Laplacian of Gaussian operator (LoG). An iterative algorithm is needed to find stable interest points, where both location and scale converge. However, it is also possible to use a faster approach with small accuracy loss. More details about the Harris-Laplace detector can be found in [19].

2.2. Difference of Gaussian detector

The difference of Gaussian (DoG) detector looks for local maxima and minima in scale space. The original image is convolved with Gaussians filters at different scales, and extrema points (over the x, y and scale directions) are selected as candidate keypoints. Equation 3 represents as $L(x, y, \sigma_i)$ the original image convolved with a Gaussian of scale σ_i, $D(x, y, \sigma_i, \sigma_j)$ being the DoG image where extrema points are selected.

$$D(x, y, \sigma_1, \sigma_2) = L(x, y, \sigma_1) - L(x, y, \sigma_2) \qquad (3)$$

A further step rejects unstable keypoints by computing location, scale and ratio of principal curvatures for each candidate and discarding those with low contrast or poorly localized. More details about the DoG interest point detector can be found in [17].

3. Scale Invariant Feature Descriptor (SIFT)

In our proposal, every interest point detected either with the Harris-Laplace detector or the DoG detector is described by means of the SIFT (Scale Invariant Feature Transform) descriptor. Basically, the neighborhood of the interest point is described with a set of orientation histograms. The result is invariant to scale, rotation (as every detected point has an associated scale σ and an associated orientation) and also to lighting, and (partially) to viewpoint change. The most common SIFT descriptor (the one that has been used in our paper) has 128 dimensions, which correspond to 16 histograms of 8 bins each computed around the pixel neighborhood. The SIFT descriptor is fully detailed in [17].

4. Proposed approach

4.1. Interest point detection in faces

Both the Harris-Laplace detector and the DoG detector are used, in order to obtain as many interest points as possible in each image. The results obtained by both detectors are shown in figure 1, when applied to one of the images of the AT&T database of faces. It becomes clear that each detector looks for specific features (blob-like for DoG and corner-like for Harris-Laplace), so combining both sets of points seems to be helpful for describing each face.

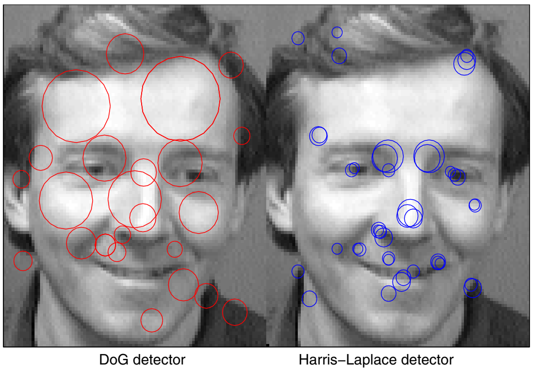

Figure 1. Comparison of interest point detectors on a face image (AT&T database).

4.2. Distance computation

Our main goal is to avoid the use of the euclidean distance between matching descriptors found in two images as a measure of the difference between such images. Similar approaches have been proposed by many authors, but in our opinion they are not completely reliable, as the absolute value of the distance between two SIFT descriptors is not proportional to the quality of the match, according to Lowe's work [17]. According to this author, the ratio between the distances of the best and the second best matches is much more reliable.

We use such an strategy in order to establish a distance measure between two images. For each descriptor found in the first image, we compute the ratio between the best and second best matches for the second image, and we count the number of matches whose ratio falls over a fixed threshold. Such a number is a first measure of distance between images (the higher the number, the closer the images). However, some false matches may fulfil the previous condition, so there is a need for a further refinement of the results.

In order to perform such refinement, Lowe's proposals for object recognition using SIFT descriptors are also taken into account. As a second measure of distance between images, we consider the number of matching points that fulfil an extra condition: they are coherent in terms of scale and orientation. Both detectors used (Harris-Laplace and DoG) associate a certain scale and orientation to each detected point (scale and orientation where a certain property is maximal). The ratio between the scales of the matching points in the two images must be approximately constant if the matches are correct, as well as the difference between the orientations. A Hough transform is used to select the most coherent matches. The number of matches fulfilling this new condition is the second measure of image similarity.

Finally, an extra filtering is performed in order to discard possible false matches, following Lowe's proposal again. By means of a second Hough transform, only those matching points whose relative location is coherent are kept. Although not completely correct, as faces are non-planar, an affine transform is considered and the parameters of such transform are required to be roughly constant among all the matching points. The number of matching descriptors fulfilling this new requirement is our third measure of image similarity.

These three measures could be used as an input to a classifier, like a Support Vector Machine, but in our paper we follow a simpler approach: we establish fixed weights for each measure, thus obtaining a single similarity measure combining all the information. Equation 4 shows as S^{a-b} the similarity between images a and b, M_D being the total number of matches between descriptors, M_{SO} being the matches fulfilling scale and orientation coherence, and M_{RL} being the matches fulfilling relative orientation coherence.

$$S^{a-b} = M_D^{a-b} + 10 \cdot M_{SO}^{a-b} + 100 \cdot M_{RL}^{a-b} \qquad (4)$$

As the third measure is the more reliable, it is given the highest weight and, in the same way, the second measure is given a higher weight that the first one. The difference in weights has been arbitrarily set to be exactly one order of magnitude. The reason why the three measures are used (and not only the most reliable one) is that face images of the same person may present a small number of matching descriptors, particularly when there are severe occlusions, and M_{RL} or even M_{SO} may turn out to be zero. By using the three measures our algorithm becomes more robust under such circumstances.

The process of obtaining M_{SO} and M_{RL} from M_D is shown over an example face image in figure 2. The first row shows the complete set of matching descriptors, where some false matches can be clearly seen. The second and third rows show reduced sets of matching points, those fulfilling respectively scale+orientation and relative location coherence. False matches are discarded, at the expense of discarding also some of the correct matches.

Once the distance between two images has been defined, a simple nearest neighbor approach is used to classify new faces: the class assigned is that of the nearest training example.

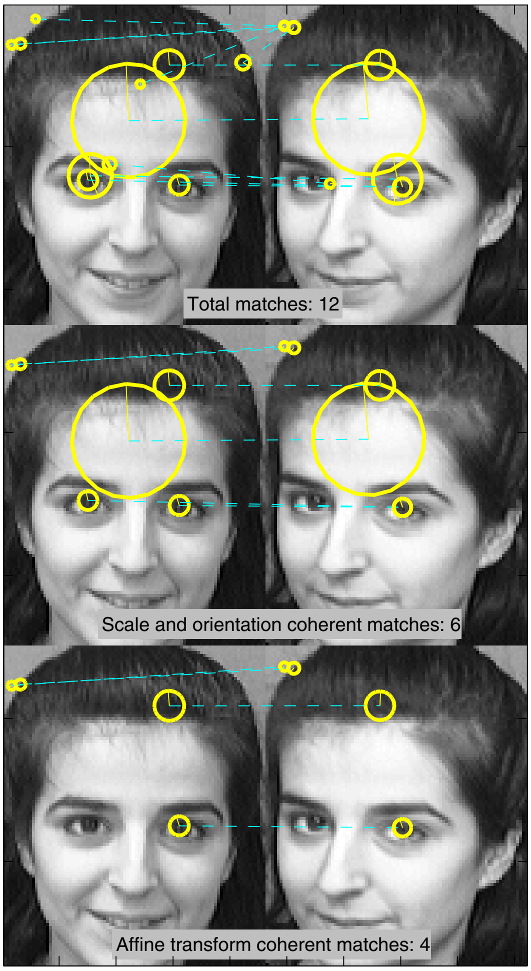
Total matches: 12
Scale and orientation coherent matches: 6
Affine transform coherent matches: 4

Figure 2. First row: all matching descriptors; second row: only those coherent to scale and orientation; third row: only those coherent to relative location.

5. Comparison with classic approaches

5.1. Databases used for the comparison

Two databases have been used for our experiments: the well known AT&T database [22], containing 400 images (40 subjects) and our own database, containing 510 images (17 subjects), which we can supply for research purposes upon request. We will refer to this database as UMH database, as the subjects are students from University Miguel Hernandez.

Concerning the AT&T database, all images were taken against a dark, homogeneous background. The main variations between the 10 shots of each subject were caused by different facial expressions, different lighting and the presence or absence of glasses. An extra source of variation comes from the viewpoint change, as the subjects are not always perfectly facing the camera.

Concerning the UMH database, there are 30 images per subject. Among these 30 images, there are 3 different groups: images #1 to #10 are taken against a uniform, white background; images #11 to #20 are taken against a non-uniform background; finally, images #21 to #30 are the most challenging ones, as part of the face is hidden by sunglasses or scarfs. Figure 3 shows some example images of one subject; rows 1 to 3 correspond to image groups 1 to 3.

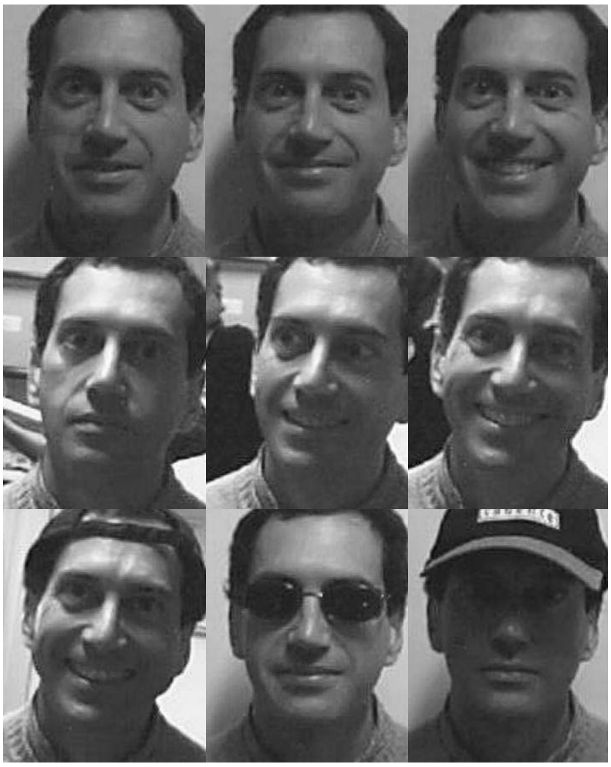

Figure 3. Some images of the UMH database of faces; each row corresponds to a more challenging set of shots.

5.2. PCA and ICA for face recognition

PCA (principal component analysis), ICA (independent component analysis) and their variations are among the most widely used algorithms for face recognition. We used them as a baseline for establishing the performance of our approach.

First, we performed some tests in order to find the best parameter settings (namely, the number of components) and the best method (among PCA, whitened PCA and ICA) for our face recognition problem (in [26] some of these feature extraction methods are shown to be equivalent under certain restrictions, such as the use of rotational invariant classifiers).

Using the two image databases available, we followed a leave-one-out approach in order to check whether a certain image of a person was classified as belonging to such person or not. The results of our parameter adjustment tests are shown in figures 4 (AT&T database) and 5 (UMH database).

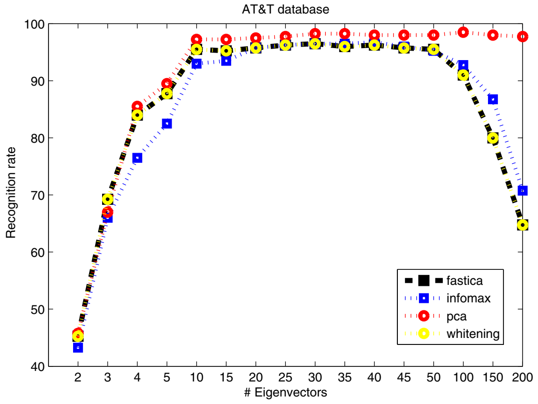

Figure 4. PCA/ICA results for the AT&T database.

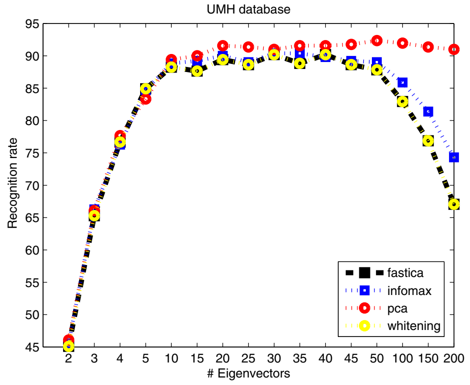

Figure 5. PCA/ICA results for the UMH database.

According to the results obtained with both databases, ICA (either using the FastICA implementation [10] or the Infomax implementation [5]) does not improve the results obtained by PCA; and whitened PCA offers exactly the same results as FastICA (as predicted in [26]), thus not improving the PCA results. Being the simpler method and offering the best results, PCA was chosen for our comparison tests. Concerning the number of components, the results with the AT&T database show that the PCA classification accuracy is approximately constant from 20 components on, with a small peak at 100 components; while the results with the UMH database show a similar behavior, with the peak at 50 components (in both cases, the performance of ICA or whitened PCA start to degrade when the number of components is that high). In order to use the best possible baseline, 100 components were used for the AT&T database and 50 components were used for the UMH database.

Database	AT&T	UMH
Images/subjects	400/40	510/17
PCA correctly classified faces	394	471
SIFT correctly classified faces	397	481
PCA classification accuracy	98.50%	92.35%
SIFT classification accuracy	99.25%	94.31%

Table 1. Comparison of classification accuracy.

5.3. Comparative analysis

Both the recognition and the authentication performance of our algorithm were compared to those of PCA. In a face *recognition* scenario, the goal is to compare a new face image to all the images in a database, in order to identify the subject (possible applications include searches in criminal databases). We measured the recognition performance in terms of classification accuracy. On the other side, in an *authentication* scenario, the goal is to confirm whether a claimed identity is true (possible applications include access control to buildings or airports). We measured the authentication performance with the AUC (area under curve) of ROC curves (receiver operation characteristic curves).

Concerning the recognition performance, we performed two different tests. First, we followed a leave-one-out strategy, and our algorithm outperformed the PCA approach on both databases. Table 1 shows the results obtained. The recognition rates were as high as 99.25% for the AT&T database (397 correct classifications from 400 images) and 94.31% for the UMH database (481 correct classifications from 510 images, including 170 images with uncontrolled backgrounds and 170 images with occlusions).

However, a leave-one-out strategy is not too realistic, as the availability of training examples is much more limited in real applications. In order to obtain results under more realistic conditions, we fixed the number of training examples to a percentage of the total images ranging from 10% to 90% (the remaining images were used as test examples). Reliability of the results was assured by performing 50 repetitions for each percentage, with different random selections of the training examples. Figures 6 and 7 show the results obtained for the AT&T database and the UMH database, respectively (mean values and standard deviations are displayed). For the AT&T database, the results offered by both algorithms are comparable, although our algorithm offers slightly better recognition rates. For the UMH database, the superiority of our approach is more evident, and increases with the percentage of test examples (increases as the recognition problem becomes more challenging).

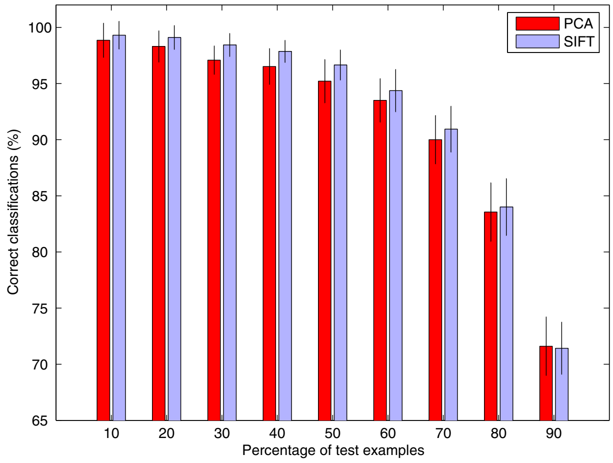

Figure 6. Results with different percentages of test examples (AT&T database).

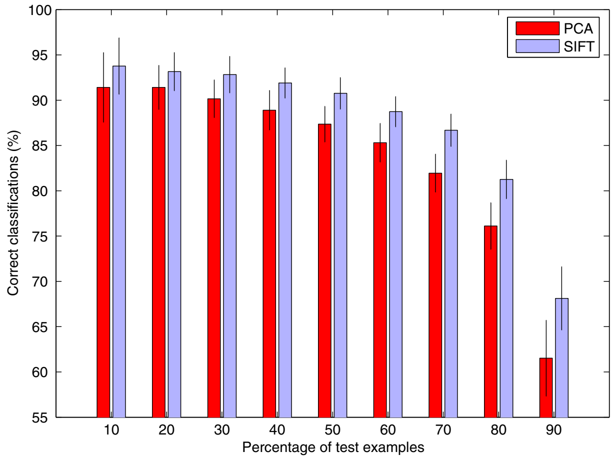

Figure 7. Results with different percentages of test examples (UMH database).

Concerning the authentication performance, ROC curves have been obtained for every subject of both databases. Again, our algorithm outperforms the PCA approach. Figure 8 shows the AUC mean values obtained for every subject of the AT&T database. Our approach gives ideal ROC curves (100% area under curve) for 34 out of 40 subjects, and it performs better than PCA for 11 subjects, while the PCA approach performs better for only 3 subjects. Obtaining the mean over all subjects, the superiority of our algorithm is also clear. Figure 9 shows the results obtained with the UMH database. Despite the authentication being more challenging due to the complexity of the database, our algorithm obtains ideal ROC curves for 4 of the subjects, and outperforms PCA for 11 subjects (PCA outperforms our method for 5 subjects). The AUC mean value over all subjects is also higher for our method.

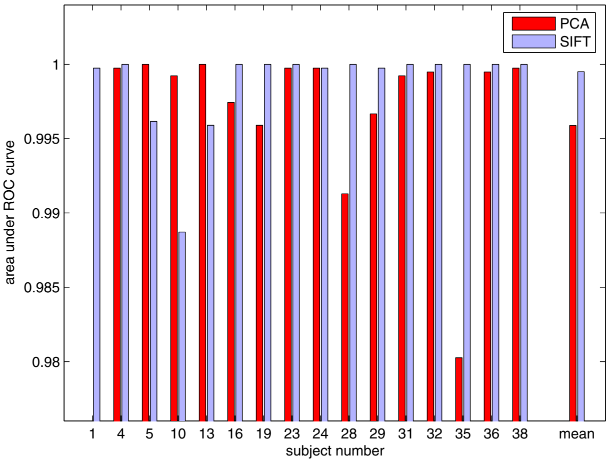

Figure 8. AUC comparison (AT&T database). Note: subjects with 100% area under curve are not plotted and PCA value for subject #1 falls out of scale (0.877).

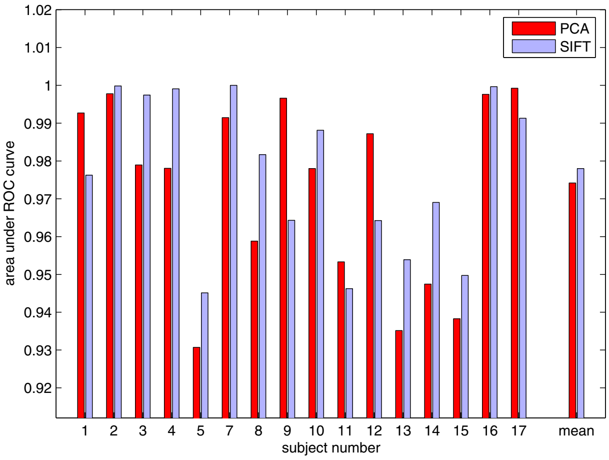

Figure 9. AUC comparison (UMH database).

6. Conclusions

It has been shown that a model-based approach using SIFT descriptors and two interest point detectors (Harris-Laplace and DoG) is valid for face recognition and face authentication.

Concerning the measure of similarity among images, the absolute value of the difference between descriptors should not be used; the number of matching descriptors should be used instead.

A simple similarity measure weighting the number of matches coherent under different restrictions (scale, orientation and relative location) is enough to obtain better results than those of classic appearance-based approaches.

As an additional conclusion, it has been shown that, among the appearance based approaches, ICA or whitened PCA do not improve the results of PCA in our face recognition scenario, where a rotational invariant classifier (namely, nearest neighbor) is used.

Future work includes code optimization and computing time measurements (at present non-optimized Matlab code is used).

7. Acknowledgments

The work presented in this paper has been funded by Generalitat Valenciana, in the frame of the project GV07/292.

References

[1] L. Alvarez and F. Morales. Affine morphological multiscale analysis of corners and multiple junctions. *Int. J. Computer Vision*, 2(25):95–107, 1997. 2

[2] M. S. Bartlett, J. R. Movellan, and J. Sejnowski. Face recognition by independent component analysis. *IEEE. Trans. Neural Networks*, 13(6):1450–1464, 2002. 1

[3] A. Baumberg. Reliable feature matching across widely separated views. *Proc. Conf. on Computer Vision and Pattern Recognition*, pages 774–781, 2000. 2

[4] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. *Proc. 4th European Conf. on Computer Vision*, pages 45–58, 1996. 1

[5] A. J. Bell and T. J. Sejnowsky. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7:1129–1159, 1995. 5

[6] M. Bicego, A. Lagorio, E. Grosso, and M. Tistarelli. On the use of SIFT features for face authentication. *Proc. Conf. on Computer Vision and Pattern Recognition Workshop*, page 35, 2006. 1

[7] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 23(6):681–685, 2001. 1

[8] B. Draper, K. Baek, M. S. Bartlett, and R. Beveridge. Recognizing faces with PCA and ICA. *Computer Vision and Image Understanding*, 91(1):115–137, 2003. 1

[9] C. Harris and M. Stephens. A combined corner and edge detector. *Proc. Alvey Vision Conference*, pages 147–151, 1998. 2

[10] A. Hyvarinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. Neural Networks*, 10(3):626–634, 1999. 5

[11] M. J. Jones and T. Poggio. Multidimensional morphable models: A framework for representing and matching object classes. *Int. J. Computer Vision*, 2(29):107–131, 1998. 1

[12] M. Kirby and L. Sirovich. Application of the Karhunen-Loeve procedure for the characterization of human faces. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 12(1):103–108, 1990. 1

[13] A. Lanitis, C. J. Taylor, and T. F. Cootes. Automatic interpretation and coding of face images using flexible models. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 19(7):743–756, 1997. 1

[14] C. Liu. Enhanced independent component analysis and its application to content based face image retrieval. *IEEE Trans. Systems, Man, and Cybernetics, Part B: Cybernetics*, 34(2):117–1127, 2004. 1

[15] D. G. Lowe. Object recognition from local scale-invariant features. *Proc. Int. Conf. on Computer Vision*, pages 1150–1157, 1999. 1

[16] D. G. Lowe. Towards a computational model for object recognition in IT cortex. *Lecture Notes in Computer Science*, 1811:20–31, 2000. 1

[17] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Computer Vision*, 60(2):91–110, 2004. 1, 2, 3

[18] S. Lucey, I. Matthews, C. Hu, Z. Ambadar, F. D. la Torre, and J. F. Cohn. AMM derived face representations for robust facial action recognition. *Proc. IEEE Int. Conf. on Automatic Face and Gesture Recognition,*, pages 155–160, 2006. 1

[19] K. Mikolajczyk and C. Schmid. Scale and affine invariant interest point detectors. *Int. J. Computer Vision*, 60(1):63–86, 2004. 2

[20] D. Ozkan and P. Duygulu. A graph based approach for naming faces in news photos. *Computer Vision and Pattern Recognition*, 2:1477–1482, 2006. 1

[21] T. T. Pham, N. E. Maillot, J. H. Lim, and J. P. Chevallet. Latent semantic fusion model for image retrieval and annotation. *Proc. 16th ACM Conf. on Information and Knowledge Management*, pages 439–444, 2007. 1

[22] F. Samaria and A. Harter. Parameterisation of a stochastic model for human face identification. *Proc. IEEE Workshop on Applications of Computer Vision*, pages 138–142, 1994. 4

[23] S. Se, D. G. Lowe, and J. Little. Vision-based mobile robot localization and mapping using scale-invariant features. *Proc. IEEE Int. Conf. on Robotics and Automation*, pages 2051–2058, 2001. 1

[24] J. Sivic, M. Everingham, and A. Zisserman. Person spotting: Video shot retrieval for face sets. *Lecture Notes in Computer Science*, 3568:226–236, 2005. 1

[25] M. Turk and A. Pentland. Eigenfaces for recognition. *J. Cognitive Neuroscience*, 3(1):71–86, 1991. 1

[26] M. A. Vicente, P. O. Hoyer, and A. Hyvrinen. Equivalence of some common linear feature extraction techniques for appearance-based object recognition tasks. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 29(5):896–900, 2007. 4, 5

[27] W. Zhao, R. Chellappa, A. Rosenfeld, and P. J. Phillips. Face recognition: A literature survey. *ACM Computing Surveys*, pages 399–458, 2003. 1