Independent Comparative Study of PCA, ICA, and LDA on the FERET Data Set

Kresimir Delac, Mislav Grgic, Sonja Grgic

University of Zagreb, FER, Unska 3/XII, Zagreb, Croatia

Received 28 December 2004; accepted 27 February 2006

ABSTRACT: Face recognition is one of the most successful applications of image analysis and understanding and has gained much attention in recent years. Various algorithms were proposed and research groups across the world reported different and often contradictory results when comparing them. The aim of this paper is to present an independent, comparative study of three most popular appearance-based face recognition projection methods (PCA, ICA, and LDA) in completely equal working conditions regarding preprocessing and algorithm implementation. We are motivated by the lack of direct and detailed independent comparisons of all possible algorithm implementations (e.g., all projection–metric combinations) in available literature. For consistency with other studies, FERET data set is used with its standard tests (gallery and probe sets). Our results show that no particular projection–metric combination is the best across all standard FERET tests and the choice of appropriate projection–metric combination can only be made for a specific task. Our results are compared to other available studies and some discrepancies are pointed out. As an additional contribution, we also introduce our new idea of hypothesis testing across all ranks when comparing performance results. © 2006 Wiley Periodicals, Inc. Int J Imaging Syst Technol, 15, 252–260, 2005; Published online in Wiley InterScience (www.interscience.wiley.com). DOI 10.1002/ima.20059

Key words: face recognition; PCA; ICA; LDA; FERET; subspace analysis methods

I. INTRODUCTION

Over the last ten years or so, face recognition has become a popular area of research in computer vision and one of the most successful applications of image analysis and understanding. Because of the nature of the problem, not only computer science researchers are interested in it, but also neuroscientists and psychologists. It is the general opinion that advances in computer vision research will provide useful insights to neuroscientists and psychologists into how human brain works, and vice versa. A general statement of the face recognition problem can be formulated as follows (Zhao et al., 2003): Given still or video images of a scene, identify or verify one or more persons in the scene using a stored database of faces. A survey of face recognition techniques has been given by Zhao et al., (2003). In general, face recognition techniques can be divided into two groups based on the face representation they use:

1. *Appearance-based*, which uses holistic texture features and is applied to either whole-face or specific regions in a face image;
2. *Feature-based*, which uses geometric facial features (mouth, eyes, brows, cheeks etc.) and geometric relationships between them.

Among many approaches to the problem of face recognition, appearance-based subspace analysis, although one of the oldest, still gives the most promising results. Subspace analysis is done by projecting an image into a lower dimensional space (subspace) and after that recognition is performed by measuring the distances between known images and the image to be recognized. The most challenging part of such a system is finding an adequate subspace.

In this paper, three most popular appearance-based subspace projection methods for face recognition will be presented, and they will be combined with four common distance metrics. Projection methods to be presented are: Principal Component Analysis (PCA), Independent Component Analysis (ICA), and Linear Discriminant Analysis (LDA). PCA (Turk and Pentland, 1991) finds a set of the most representative projection vectors such that the projected samples retain most information about original samples. ICA (Bartlett et al., 2002; Draper et al., 2003) captures both second and higher-order statistics and projects the input data onto the basis vectors that are as statistically independent as possible. LDA (Belhumeur et al., 1996; Zhao et al., 1998) uses the class information and finds a set of vectors that maximize the between-class scatter while minimizing the within-class scatter. Distance metrics used are L1 (City block), L2 (Euclidean), cosine and Mahalanobis distance.

The aim of this paper is to provide an independent, comparative study of these three projection methods and their accompanied distance metrics in completely equal working conditions. In order to perform a fair comparison, same preprocessed images are the input into all algorithms and the number of dimensions to be retained is chosen following the standard recommendations. For consistency

Correspondence to: K. Delac; E-mail: kdelac@ieee.org

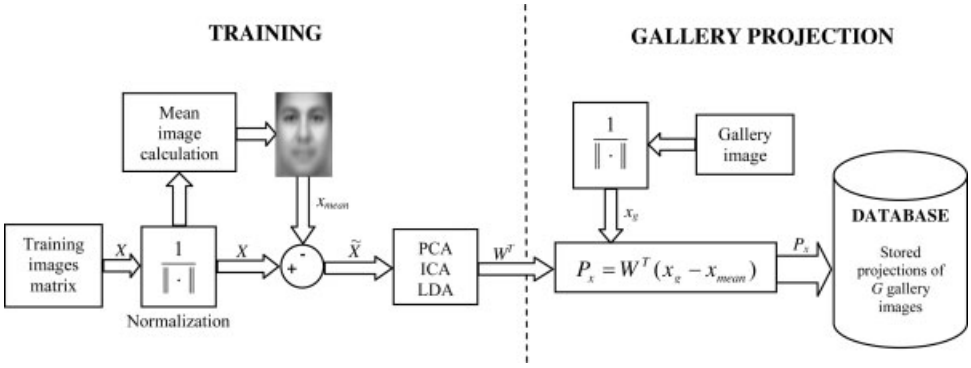

Figure 1. An illustration of general subspace appearance-based face recognition system.

with other studies, FERET data set (Phillips et al., 2000), with its standard test sets, is used for comparisons. This research is motivated by the lack of direct and detailed comparisons of these three projection methods. They are rarely compared in a single paper and almost never are all possible implementations considered (e.g., all projection–metric combinations). It is interesting to notice that the findings of other research groups are often contradictory on this subject and this is another important reason for performing a study of this kind. For example, Liu and Wechsler (1999) and Bartlett et al. (2002) claim that ICA outperforms PCA, while Baek et al. (2002) claim that PCA is better. Moghaddam (2002) states that there is no significant difference. Beveridge et al. (2001a) claim that in their tests LDA performed uniformly worse than PCA, Martinez and Kak (2001) state that LDA is better for some tasks, and Belhumeur et al. (1996) and Navarrete and Ruiz-del-Solar (2002) claim that LDA outperforms PCA on all tasks in their tests (for more than two samples per class in training phase). All these results are in most cases given only for one or two projection–metric combinations for a specific projection method, and in some cases using nonstandard databases or some hybrid test sets derived from a standard database.

The rest of this paper is organized as follows: Section II gives a brief description of the algorithms to be compared, Section III reports the details of methodology, Section IV presents the results and compares our results to results of other research groups and Section V concludes the paper.

II. ALGORITHMS

Even though projection methods and metrics used in this work are already well known, we will include a brief description for the sake of completeness. All three projection methods are so called *subspace analysis methods*. A 2D image Γ with m rows and n columns can be viewed as a vector (after concatenating its rows or columns) in N dimensional image space ($\Re^{N=m\times n}$). Since space derived this way is highly dimensional, recognition in it is unfeasible. Therefore, recognition algorithms usually derive lower dimensional spaces to do the actual recognition while retaining as much information (energy) from the original images as possible. We will further clarify this on the example from this research: the original FERET images (after preprocessing) are the size of 60 × 50 pixels, thus the image space dimensionality is $\Re^{N=60\times 50=3000}$. It will be shown that projection methods presented here will yield $\Re^{270}$ ($\Re^{224}$ for LDA) subspace in which the recognition will be done and in these 270 dimensions 97.85% of original information (energy) is retained. An example of building a general subspace appearance-based face recognition system can be seen in Figure 1. Training of the subspace system can be seen in the left part of the figure and the procedure for projecting gallery images onto a subspace (projection matrix W^T) can be seen in the right part of the figure; X is a matrix containing the images expressed as vectors in its columns, x_{mean} – mean image (as a vector), $\tilde{X}$ – matrix containing mean-subtracted images in its columns, W^T – projection matrix, x_g – gallery image (as a vector). During the training phase, the projection matrix (containing the basis vectors of the subspace) is calculated and then the gallery images (the images of known persons) are projected onto that subspace and their projections are stored in a database. Later, in the matching phase (Fig. 2), new image is normalized, mean-subtracted, projected onto the same subspace as the gallery image was and its projection is then compared to stored gallery projections (the *nearest neighbor* is determined by calculating the distances d from a probe image projection to all gallery images projections and then choosing the minimum distance as a similarity measure). The identity of the most similar gallery image is then chosen to be the

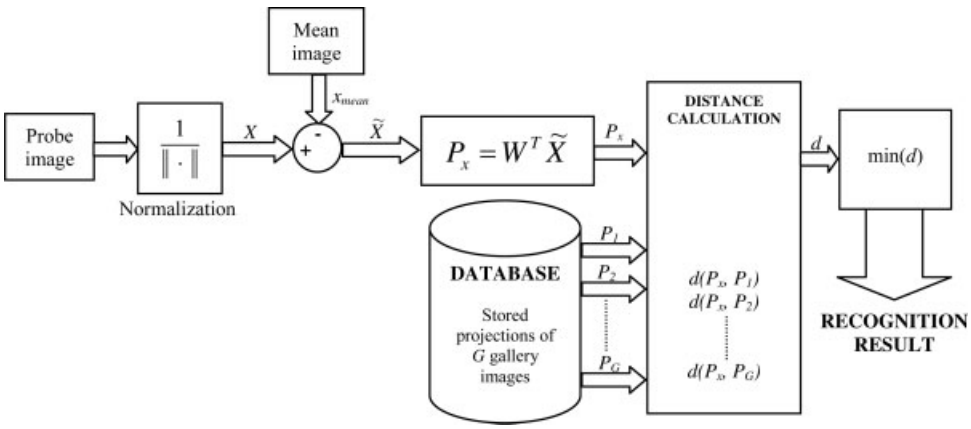

Figure 2. The matching phase of a general subspace face recognition system.

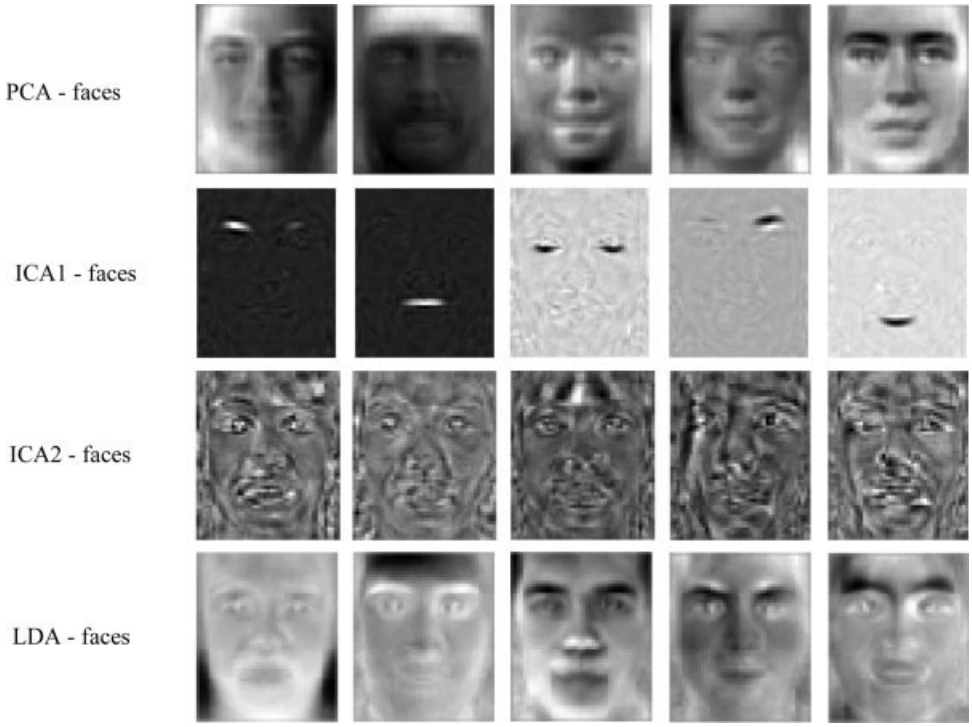

Figure 3. Face representations found by PCA (eigenfaces), ICA1, ICA2, and LDA.

result of recognition and the unknown probe image is identified. It is important to mention that a general face recognition system can work in two modes: (1) the *identification mode* where the input to the system is an unknown face and the system reports back the determined identity (our case) and (2) the *verification mode* where the system needs to confirm or reject the claimed identity of the input face. All our experiments are conducted for the identification mode and the general illustration of the systems shown in Figures 1 and 2 illustrates our experiments.

A. Principal Component Analysis (PCA). In our experiments we implemented Principal Component Analysis (PCA) procedure as described by Turk and Pentland (1991). Given an s-dimensional vector representation of each face in a training set of M images, PCA tends to find a t-dimensional subspace whose basis vectors correspond to the maximum variance direction in the original image space. This new subspace is normally lower dimensional ($t \ll s$). New basis vectors define a subspace of face images called *face space*. All images of known faces are projected onto the face space to find sets of weights that describe the contribution of each vector. To identify an unknown image, that image is projected onto the face space as well to obtain its set of weights. By comparing a set of weights for the unknown face to sets of weights of known faces, the face can be identified. If the image elements are considered as random variables, the PCA basis vectors are defined as eigenvectors of the scatter matrix S_T defined as:

$$S_T = \sum_{i=1}^{M} (x_i - \mu) \cdot (x_i - \mu)^T \tag{1}$$

where μ is the mean of all images in the training set (the *mean face*, Fig. 1) and x_i is the ith image with its columns concatenated in a vector. The projection matrix W_{PCA} is composed of t eigenvectors corresponding to t largest eigenvalues, thus creating a t-dimensional face space. Since these eigenvectors (PCA basis vectors) look like some ghostly faces they were conveniently named *eigenfaces* (Fig. 3).

B. Independent Component Analysis (ICA). PCA considered image elements as random variables with Gaussian distribution and minimized second-order statistics. Clearly, for any non-Gaussian distribution, largest variances would not correspond to PCA basis vectors. Independent Component Analysis (ICA) (Bartlett et al., 2002; Draper et al., 2003) minimizes both second-order and higher-order dependencies in the input data and attempts to find the basis along which the data (when projected onto them) are *statistically independent*. Bartlett et al. (2002) provided two architectures of ICA for face recognition task: *Architecture I* – statistically independent basis images (ICA1 in our experiments) and *Architecture II* – factorial code representation (ICA2 in our experiments).

Our implementation of ICA uses the INFOMAX algorithm proposed by Bell and Sejnowski and used by Bartlett et al. (2002). PCA is used to reduce dimensionality prior to performing ICA.

C. Linear Discriminant Analysis (LDA). Linear Discriminant Analysis (LDA) (Belhumeur et al., 1996; Zhao et al., 1998) finds the vectors in the underlying space that best discriminate among classes. For all samples of all classes the between-class scatter matrix S_B and the within-class scatter matrix S_W are defined by:

$$S_B = \sum_{i=1}^{c} M_i \cdot (x_i - \mu) \cdot (x_i - \mu)^T \tag{2}$$

$$S_W = \sum_{i=1}^{c} \sum_{x_k \epsilon X_i} (x_k - \mu_i) \cdot (x_k - \mu_i)^T \tag{3}$$

where M_i is the number of training samples in class i, c is the number of distinct classes, μ_i is the mean vector of samples belonging to class i and X_i represents the set of samples belonging to class i with x_k being the k-th image of that class. S_W represents the scatter of features around the mean of each face class and S_B represents the scatter of features around the overall mean for all face classes.

The goal is to maximize S_B while minimizing S_W, in other words, maximize the ratio $\det|S_B|/\det|S_w|$. This ratio is maximized when the column vectors of the projection matrix (W_{LDA}) are the eigenvectors of ${S_W}^{-1} \cdot S_B$. In order to prevent S_W to become singular, PCA is used as a preprocessing step and the final transformation is ${W_{opt}}^T = {W_{LDA}}^T {W_{PCA}}^T$.

In Figure 3 PCA-faces (*eigenfaces*), *ICA1-faces*, *ICA2-faces,* and *LDA-faces* can be seen. These *ghostly faces* are basis vectors produced by projection methods, reshaped to a matrix form of the original image size for convenience. This is a good illustration of the differences between subspaces derived by each of those projection methods. If we take a closer look at the basis vector representations it can be seen that PCA, LDA, and ICA2 produce global features; every image feature is influenced by every pixel. ICA1 produces spatially localized features that are only influenced by small parts of an image, thus isolating particular parts of faces. Logical conclusion is that ICA1 should be optimal for recognizing facial actions and suboptimal for recognizing temporal changes in faces or images taken under different illumination conditions. This theoretical property of ICA1 will be proven by our experiments.

D. Distance Measures. Four different distance measures will be used in comparisons: L1, L2, cosine angle (COS), and Mahalanobis distance (MAH). Generally, for two vectors, x and y, distance measures are defined as:

$$d_{L1}(x,y) = |x-y| \tag{4}$$

$$d_{L2}(x,y) = \|x-y\|^2 \tag{5}$$

$$d_{\cos}(x,y) = -\frac{x \cdot y}{\|x\| \cdot \|y\|} \tag{6}$$

$$d_{\mathrm{MAH}}(x,y) = \sqrt{(x-y)V^{-1}(x-y)^T} \tag{7}$$

where V is the covariance matrix. In the rest of this paper, we will address the specific projection–metric combination as an *algorithm*. Since we implemented four projection methods (PCA, ICA1, ICA2, and LDA) and four distance measures (L1, L2, COS, and MAH), it can be concluded that we will effectively compare 16 different algorithms.

III. METHODOLOGY

A. Data. For consistency with other studies, we used the standard FERET data set including the data partitions (subsets) for recognition tests, as described in Phillips et al., 2000. The *gallery* consists of 1,196 images and there are four sets of probe images (*fb*, *fc*, *dup1*, and *dup2*) that are compared to the *gallery* images in recognition stage. The *fb* probe set contains 1,195 images of subjects taken at the same time as *gallery* images with one difference being that the subjects were told to assume a different facial expression. The *fc* probe set contains 194 images of subjects under different illumination conditions. The *dup1* (duplicate I) set contains 722 images taken anywhere between one minute and 1,031 days after the *gallery* image was taken, and *dup2* (duplicate II) set is a subset of *dup1* containing 234 images taken at least 18 months after the *gallery* image was taken. All images in the data set are of size 384 × 256 pixels and grayscale.

B. Preprocessing. All algorithms and all image preprocessing steps were implemented in MATLAB®. Original FERET images were first spatially transformed (to get eyes at fixed points in imagery) based upon a ground truth file of eye coordinates supplied with the original FERET data. The standard *imrotate* MATLAB® function was used with bilinear interpolation parameter. After that, all images were cropped the same way (using the eyes coordinates) to eliminate as much background as possible. No masking was done since it turned out that cropping eliminated enough background and the whole idea of this research was not to yield the best possible recognition results but to fairly compare the algorithms. After cropping, images were additionally resized to be the size of 60 × 50 using the standard MATLAB® *imresize* function with bilinear interpolation. Finally, image pixel values were histogram equalized to the range of values from 0 to 255 using the standard *histeq* function.

C. Training. To train the PCA algorithm we used a subset of classes for which there were *exactly* three images per class. We found 225 such classes (different persons) in the FERET data set, so our training set consisted of $3 \times 225 = 675$ images ($M = 675$, $c = 225$). PCA derived, in accordance with theory, $M - 1 = 674$ meaningful eigenvectors. We adopted the FERET recommendation and kept the top 40% of those, resulting in 270-dimensional PCA subspace (40% of 674 ≈ 270). It was calculated that 97.85% of original information (energy) was retained in those 270 eigenvectors. This subspace was used for recognition as PCA face space and as input to ICA and LDA (PCA was the preprocessing dimensionality reduction step). ICA yielded two representations (ICA1 & ICA2) using the input from PCA (as in Bartlett et al., 2002). Dimensionality of both ICA representations was also 270. However, LDA yielded only 224-dimensional space since it, by theory, can produce a maximum of $c - 1$ basis vectors. All of those were kept to stay close to the dimensionality of PCA and ICA spaces and thus make comparisons as fair as possible. After all the subspaces have been derived, all images from data sets were projected onto each subspace and recognition using nearest neighbor classification with various distance measures was conducted.

IV. RESULTS

Results of our experiment can be seen in Table I and Figure 4. We used two standard ways to present the results: (1) table showing algorithm performance at rank 1 (recognition rate within the top one match), and (2) Cumulative Match Score (CMS) curve (Phillips et al., 2000), showing the cumulative results for ranks 1 and higher. One interesting thing we noticed is the discrepancy in some cases between the rank 1 results and the CMS results when answering the question which algorithm performs better. It was noticed that the metric showing the best results at rank 1 did not always yield the best results at higher ranks. Five such cases were identified (most frequently for LDA) in this experiment. This can be seen in Table I by comparing the bolded best algorithm–metric combinations for rank 1 and the right two columns showing the best combinations at higher ranks. This brings to question any analyses done by comparing the CMS curves of those projection–metric combinations that yielded the best results at rank 1. This is why we decided to show the CMS curves for those metrics that produced best results at higher ranks for a specific algorithm.

A. Conclusions Based on a Specific Task. ***fb*** (the *different expression* task). Even though ICA2+COS combination produces the best results at rank 1 (Table I), LDA+COS outperforms it from rank 6 further on (Fig. 4). Actually, ICA2+COS performs uniformly worse than other best projection–metric combinations at

Table I. Performance across four projection methods and four metrics.

	Results at Rank 1 (%)					
	Metric				CMS Results	
Projection	**L1**	**L2**	**MAH**	**COS**	Highest Curve	Same as Rank 1
			Fb			
PCA	**82.26**	*82.18*	*64.94*	*81.00*	PCA+COS	F
ICA1	*81.00*	**81.51**	*64.94*	*80.92*	ICA1+L2	T
ICA2	*64.94*	*74.31*	*64.94*	**83.85**	ICA2+COS	T
LDA	*78.08*	**82.76**	*70.88*	*81.51*	LDA+COS	F
			Fc			
PCA	**55.67**	*25.26*	*32.99*	*18.56*	PCA+L1	T
ICA1	*18.04*	*17.53*	**32.99**	*12.89*	ICA1+L1	F
ICA2	*15.98*	*44.85*	*32.99*	**64.95**	ICA2+COS	T
LDA	*26.80*	*26.80*	**41.24**	*20.62*	LDA+L2	F
			Dup1			
PCA	**36.29**	*33.52*	*25.62*	*33.52*	PCA+L1	T
ICA1	**32.55**	*31.86*	*25.62*	*32.27*	ICA1+L1	T
ICA2	*28.81*	*31.99*	*25.62*	**42.66**	ICA2+COS	T
LDA	**34.76**	*32.96*	*27.70*	*33.38*	LDA+L1	T
			Dup2			
PCA	**17.09**	*10.68*	*14.53*	*11.11*	PCA+L1	T
ICA1	*8.97*	*7.69*	**14.53**	*8.97*	ICA1+MAH	T
ICA2	*16.24*	*19.66*	*14.53*	**28.21**	ICA2+COS	T
LDA	*16.24*	*10.26*	**16.67**	*10.68*	LDA+L1	F

Left part contains the results for rank 1 and the best projection–metric combinations are bolded. Right part contains are the best CMS results obtained by determining which metric gives the highest curve for a specific projection method at a specific probe set.

higher ranks. But, it can be stated that the remaining three projection–metric combinations (PCA+COS, ICA1+L2, and LDA+COS) produce similar results and no straightforward conclusion can be drawn regarding which is the best for this specific task. It stays unclear whether the differences between recognition rates for this task are statistically significant or not (we will address this issue in Section IV(C)). ICA1 performance was comparable to PCA and LDA and certainly better than ICA2. This confirms the theoretical property of ICA1 mentioned in Section II(C).

fc (the *different illumination* task). ICA2+COS wins here at rank 1 (Table I) but PCA+L1 is much better from rank 17 on (Fig. 4). ICA1 is the worst choice for this task. Again, this is not surprising since ICA1 tends to isolate the face parts and therefore should be better at recognizing facial actions than anything else.

dup1 & dup2 (the *temporal changes* tasks). ICA2+COS is the best here at every rank (as clearly illustrated in Fig. 4) and ICA1 is the worst. L1 norm seems to produce the best results for almost all other projection methods for this task and it is surprising that it is so rarely used in comparisons.

B. Metrics Comparison. L1 gives the best results in combination with PCA across all four probe sets so it can be concluded that L2 distance metric is suboptimal for PCA (one exception being that COS outperforms L1 for the *fb* set, but statistical significance remains questionable). Following the same line of thought, it can be concluded that COS is superior to any other metric when used with ICA2. Actually, L2 is the best metric in only two combinations across all probe sets and projection methods. We found this result surprising since this was the most frequently used measure in the past. No clear conclusion can be drawn as to which metric works best with ICA1 and LDA and, at best, it can be stated that it depends on the nature of the task. This tells us that no combination of projection–metric for ICA1 and LDA are robust enough across all tasks. MAH turned out to be the most disappointing metric in our tests and more variants of MAH distance calculations should be investigated (as in Beveridge et al., 2001a). If we analyze the best results given by the CMS, the metrics ranking looks something like this: L1 – 7 best results, COS – 6, L2 – 2, MAH – 1.

C. Evaluating the Differences in Algorithm Performance (Hypothesis testing). We think that, when comparing recognition algorithms, it is important (yet often neglected) to answer the following question: *when is the difference in performance statistically significant?* Clearly, the difference in performance of 1% or 2% could be due to pure chance, while 10% or more is probably not. We made our conclusions so far based on our intuitive analysis of the data (recognition rate percentage was higher or the curve on the plot appeared higher). However, we felt the need to investigate these intuitive presumptions using standard statistical hypothesis testing techniques. Generally, there are two ways of looking at the performance difference (Jambor et al., 2002): (1) determine if the difference (as seen over the entire set of probe images) is significant, (2) when the algorithms behave differently, determine if the difference is significant. As argued by Jambor et al. (2002 the first way to evaluate performance difference fails to take the full advantage of the standard face recognition protocol, so we will focus on the second way and set our hypotheses like this: *H1:* when algorithms A and B differ on a particular image, A is more likely to correctly identify that image, *H0:* when algorithms A and B differ on a particular image, they are equally likely to identify that image correctly. In order to perform this test we recorded which of the four possible outcomes, when comparing two algorithms (*SS* – both successful, *FF* – both failed, *FS* – first one failed and the second one succeeded, *SF* – first one succeeded and the second one failed), is true for each probe image. Following the lead of Beveridge et al., 2001b; Jambor et al., 2002 we employed McNemar's Test with the

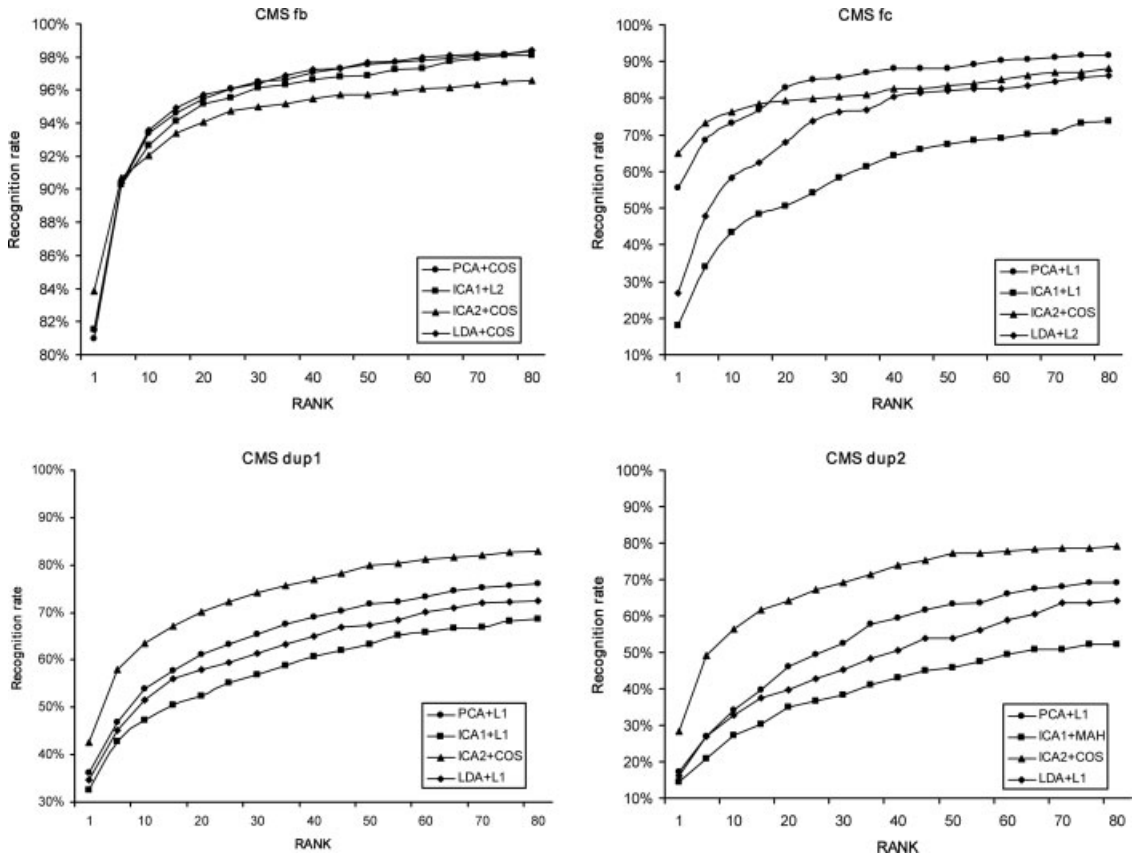

Figure 4. Cumulative Match Score (CMS curve) plots of best projection–metric combinations (the ones that yielded the highest curve when all metrics were compared for a specific algorithm) for a given probe set.

0.05 cutoff assumed. This test ignores the cases where both compared algorithms behave the same (*SS* or *FF*) and only uses the outcomes where they behave differently (*SF* and *FS*). Let n_{SF} be the number of times the *SF* is observed and n_{FS} the number of times *FS* is observed. We are interested in the one-sided version of this test so the probability of *H0* is bounded by:

$$P_{H0} \leq \sum_{i=0}^{n_{FS}} \frac{n!}{i!(n-i)!} 0.5^n \tag{8}$$

where $n = n_{SF} + n_{FS}$. In other words, *H0* will be rejected in favor of *H1* if $P_{H0} \leq 0.05$. So far, this test was usually performed only at rank 1 so we decided to expand it to all ranks and to plot the results as a step curve (Figs. 5–8).

Using the described methodology we first checked if our choice of best metric for a specific projection method was correct. We confirmed our choices of the highest curves and found that the choice was correct or, at least, that there is no significant difference between the chosen metric and the second best metric. Thus, the CMS curves compared in Figure 4 are truly the best choices of a projection–metric combination for a given task. Next, we plotted the decision graph based on McNemar's Test (Figs. 5–8) for all CMS curves given in Figure 4. All possible comparison combinations are given thus yielding six plots for every probe set.

As expected, the most complicated situation is for the *fb* probe set (Fig. 5). Here we claimed that LDA+COS outperforms ICA2+COS from rank 6 further on but this is obviously true only from rank 9 further on. For other combinations we confirmed our presumption that there is no significant difference in performance (Fig. 5).

Situation is a little less complicated for the *fc* probe set (Fig. 6). For example, we can see that there is no significant difference between ICA2+COS and LDA+L2 from rank 25 on and yet the CMS curve in Figure 4 is always higher for ICA2+COS. We also claimed that PCA+L1 is better than ICA2+COS from rank 17 on. This is not quite true since it can be seen in Figure 6 that the difference becomes significant only from rank 27 further on. However, Figure 6 confirms the statement that ICA1 is the worst choice for this task.

It can be seen that for *dup1* (Fig. 7) and *dup2* set (Fig. 8) everything is relatively clear and all our previous conclusions are confirmed by this test.

We can state that our previous overall conclusions regarding the relative performance were confirmed by this hypothesis testing technique and some new conclusions were drawn regarding the exact rank in which the differences become significant.

D. Comparison to Previous Work. It is worth mentioning at this point that most of the papers we will compare our results to do not use standard FERET test sets, but this should not be a problem since we will make the comparisons to their results based on the *relative* performance.

First of all, we can state that our results are consistent to that of Phillips et al. (2000) regarding the relative ranking of probe sets. *fb* was found to be the easiest (highest recognition rates) and *dup2* the hardest (lowest recognition rates). This is in clear

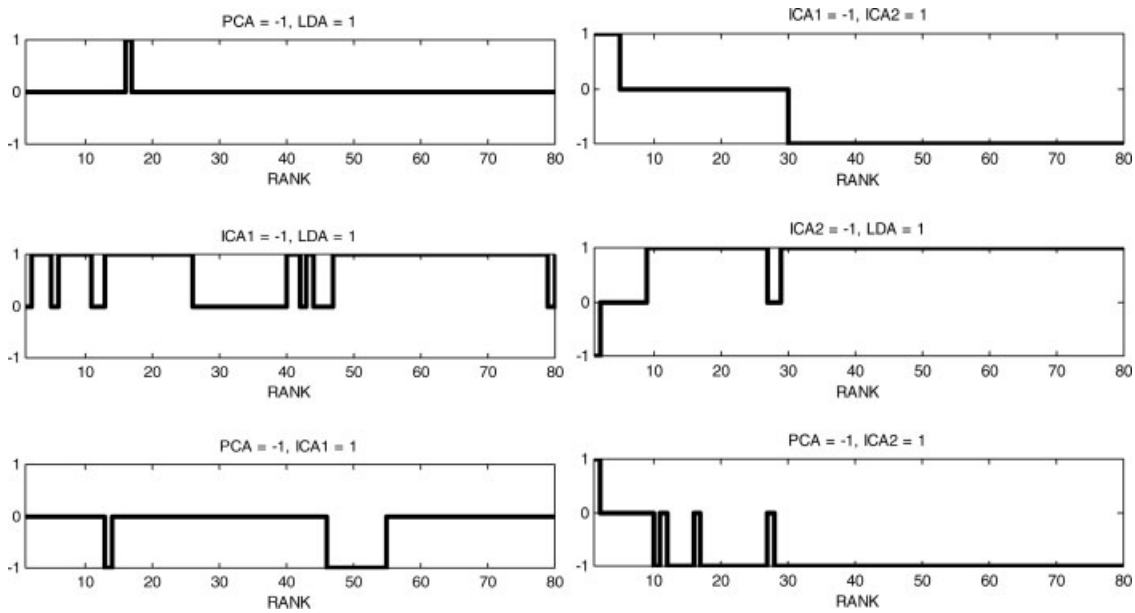

Figure 5. Hypothesis testing for the *fb* probe set—comparing the best projection–metric combinations from Figure 4. 1 or −1 means that one or the other algorithm is significantly better at a given rank and 0 means that there is no significant difference.

contradiction with Beak et al. (2002) who stated that *fc* is the hardest probe set. Also consistent with Phillips et al. (2000) is that LDA+COS outperforms all others for the *fb* set (or at least that there is no significant difference). Both Phillips et al. (2000) and Draper et al. (2003), when comparing PCA and ICA, claim that ICA2 outperforms PCA+L2 and this is what we also have found. However, our detailed research also produced some new conclusions: (1) PCA+COS outperforms ICA2+COS for *fb* probe set at higher ranks and (2) PCA+L1 outperforms ICA2−COS for *fc* probe set at higher ranks (this is also confirmed by hypothesis testing in Figs. 5–6). Those new conclusions are consistent to that of Bartlett et al. (2002) who favor ICA2 over PCA (actually, mostly on difficult time-separated images), but we disagree with their claim that ICA2 is better for the *different expression* task also. As stated by Bartlett et al. (2002), we also found that ICA2 gives best results when combined with COS. Navarrete and Ruiz-del-Solar (2002) claim that LDA+COS works better than PCA, which is certainly not the case here at rank 1 and is questionably true for higher ranks. We agree with Moghaddam (2002) who stated that there is no significant difference between PCA and ICA at rank 1, but we think that ICA is significantly worse at higher ranks.

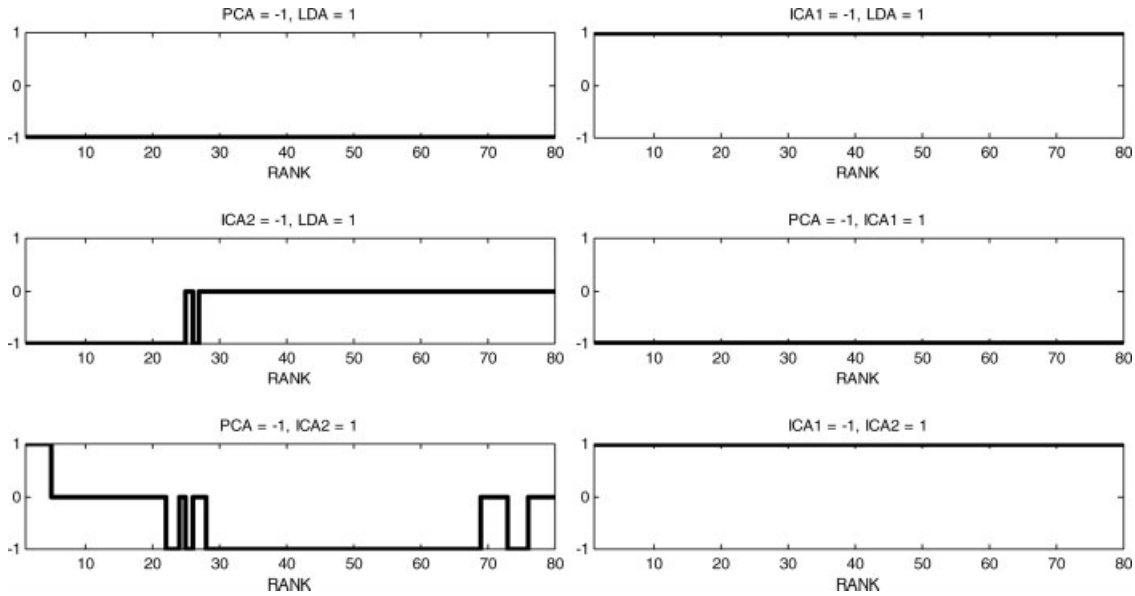

Figure 6. Hypothesis testing for the *fc* probe set—comparing the best projection–metric combinations from Figure 4. 1 or −1 means that one or the other algorithm is significantly better at a given rank and 0 means that there is no significant difference.

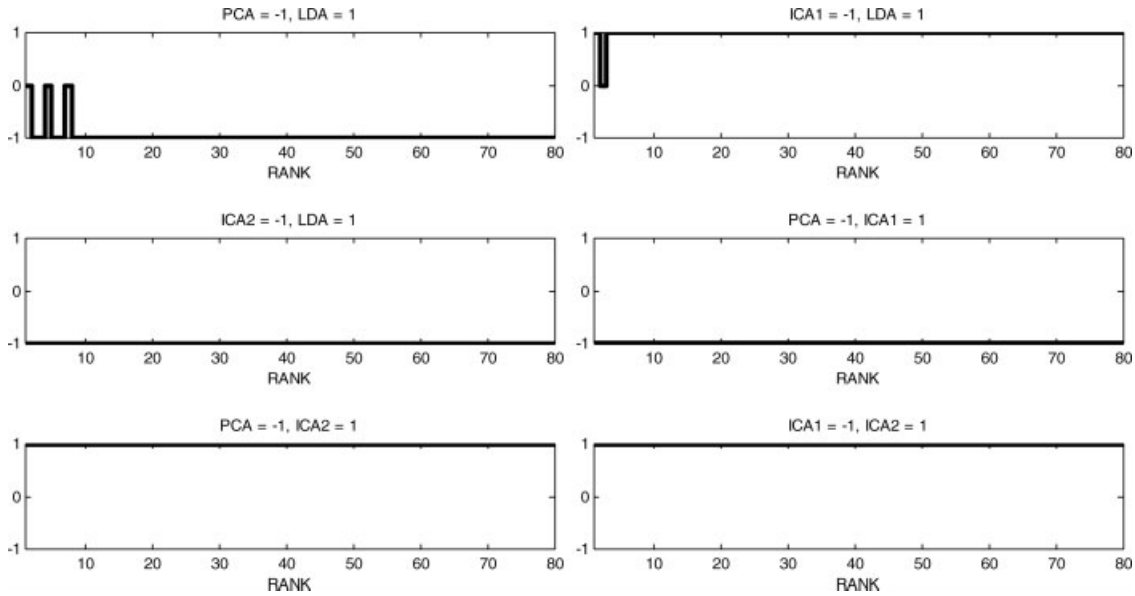

Figure 7. Hypothesis testing for the *dup1* probe set—comparing the best projection–metric combinations from Figure 4. 1 or −1 means that one or the other algorithm is significantly better at a given rank and 0 means that there is no significant difference.

Among the most surprising results was the poor performance of LDA in some cases. Contrary to FERET's report in (Phillips et al., 2000) where the LDA implementation from UMD performed consistently better than the baseline PCA (Turk and Pentland, 1991), we found that this is not true in our experiments. However, the FERET test did not measure the effect that individual algorithm's parts (components) have on overall performance and it did not concern with algorithm implementation details. One of the most important parts of design and implementation of an algorithm is the training stage. It is unclear from Phillips et al., 2000 (and also from Zhao et al., 1998) exactly how was the LDA trained. Zhao et al. (1998) it is stated that the algorithm is trained using the 1038 FERET images from 444 classes and in (Phillips et al., 2000) the baseline PCA was trained with randomly chosen 500 FERET images. These two different training scenarios make it very difficult to directly compare these two algorithms and their various components. Even if we were to compare them, it is obvious that LDA should have a great advantage as it was trained with roughly two times bigger training set. The poor performance of our implementation of LDA could also be due to the relatively small training set. Similar problem was identified and researched in detail in Martinez and Kak, 2001 where it is concluded that when the training set is small, PCA can outperform LDA. Another important thing to mention is that LDA is much more sensitive to different training sets than PCA or ICA. There is

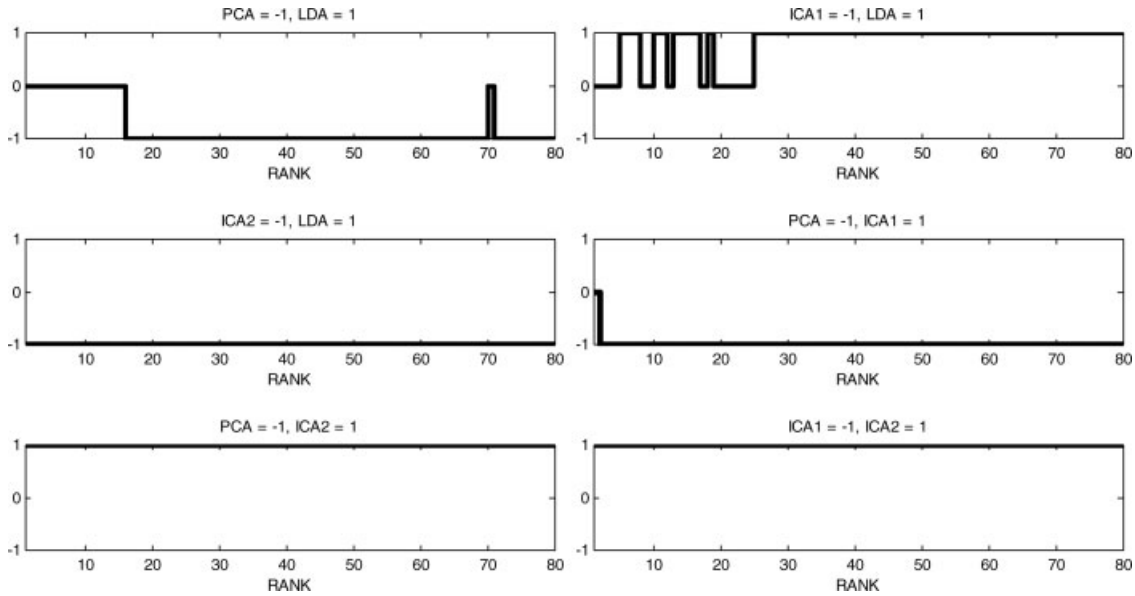

Figure 8. Hypothesis testing for the *dup2* probe set—comparing the best projection–metric combinations from Figure 4. 1 or −1 means that one or the other algorithm is significantly better at a given rank and 0 means that there is no significant difference.

reason to believe that a different training set could yield different, possibly better, results for LDA. But again, the goal of this paper was to introduce a direct comparison in equal working conditions. By using only three images per class for training we tried to stay close to real-life law enforcement applications, where it is difficult to expect to have more than a few images of an individual. It is also very important to mention that our training set and query set overlap in only 33% of classes (so some of the classes later used in tests are new to the systems), whereas many other papers report results training the algorithms with the images of same individuals later to be used in tests. This makes our tests bit more difficult for the algorithms.

V. CONCLUSION

This paper presented an independent, comparative study of three most popular appearance-based face recognition projection methods (PCA, ICA, and LDA) and their accompanied four distance metrics (L1, L2, cosine, and Mahalanobis) in completely equal working conditions. This experimental setup yielded 16 different algorithms to be compared. From our independent comparative research we can derive the following conclusions: (1) no claim can be made about which is the best combination for the *different expression* task since the differences do not seem to be statistically significant (although LDA+COS seems to be promising), (2) PCA+L1 outperforms ICA1 and LDA with *illumination changes* task at all ranks and outperforms ICA2 from rank 27 further on, (3) COS seems to be the best choice of metric for ICA2 and gives good (but not always the best compared to other projection methods) results for all probe sets, (4) ICA2+COS combination turned out to be the best choice for *temporal changes* task, (5) in many cases L2 produced lower results than L1 or COS and it is surprising that it was used so often in the past, (6) L1 and COS metrics produced best overall results across all algorithms and should be further investigated. Finally, it can be stated that, when tested in completely equal working conditions, no algorithm (projection–metric combination) can be considered the best time and the choice of appropriate algorithm can only be made for a specific task. Some theoretical properties of the described algorithms were also confirmed and illustrated.

We also presented a new methodology for comparing two CMS curves, based on McNemar's Test of statistical significance. It was shown that straightforward conclusions based only on the inspection of the two curves should not be drawn, since the difference in performance could easily be insignificant and some wrong conclusions could be made regarding the exact rank at which the curves start to differ.

ACKNOWLEDGMENTS

Portions of the research in this paper use the Color FERET database of facial images collected under the FERET program.

REFERENCES

K. Baek, B. Draper, J.R. Beveridge, and K. She, PCA vs. ICA: A comparison on the FERET data set, Proc Fourth Int Conf Computer Vision, Pattern Recogn and Image Process, Durham, NC, 8–14 March, 2002, pp. 824–827.

M.S. Bartlett, J.R. Movellan, and T.J. Sejnowski, Face recognition by independent component analysis, IEEE Trans Neural Networks 13 (2002), 1450–1464.

P. Belhumeur, J. Hespanha, and D. Kriegman, Eigenfaces vs. fisherfaces: Recognition using class specific linear projection, Proc Fourth Eur Conf Computer Vision, Vol. 1, 14–18 April 1996, Cambridge, UK, pp. 45–58.

J.R. Beveridge, K. She, B. Draper, and G.H. Givens, A nonparametric statistical comparison of principal component and linear discriminant subspaces for face recognition, Proc IEEE Conf Computer Vision and Pattern Recogn, Kauai, HI, December, 2001a, pp. 535–542.

R. Beveridge, K. She, B. Draper, and G. Givens, Parametric and non-parametric methods for the statistical evaluation of human ID algorithms, IEEE Third Workshop on Empirical Evaluation Methods in Computer Vision, Kauai, HI, December, 2001b.

B. Draper, K. Baek, M.S. Bartlett, and J.R. Beveridge, Recognizing faces with PCA and ICA, Comput Vis Image Understand (Special Issue on Face Recognition) 91 (2003), 115–137.

W.S. Jambor, B.A. Draper, and J.R. Beveridge, "Analyzing PCA-based face recognition algorithms: Eigenvector selection and distance measures," In: H. Christensen, J. Phillips (Editors), Empirical evaluation methods in computer vision, World Scientific, Singapore, 2002.

C. Liu and H. Wechsler, Comparative assessment of independent component analysis (ICA) for face recognition, Second International Conference on Audio- and Video-based Biometric Person Authentication, Washington DC, 22–23 March, 1999.

A. Martinez and A. Kak, PCA versus LDA, IEEE Trans Pattern Anal Machine Intell 23 (2001), 228–233.

B. Moghaddam, Principal manifolds and probabilistic subspaces for visual recognition, IEEE Trans Pattern Anal Machine Intell 24 (2002), 780–788.

P. Navarrete and J. Ruiz-del-Solar, Analysis and comparison of Eigenspace-based face recognition approaches, Int J Pattern Recogn Artif Intell 16 (2002), 817–830.

P.J. Phillips, H. Moon, S.A. Rizvi, and P.J. Rauss, The FERET evaluation methodology for face recognition algorithms, IEEE Trans Pattern Anal Machine Intell 22 (2000), 1090–1104.

M. Turk and A. Pentland, Eigenfaces for recognition, J Cogn Neurosci 3 (1991), 71–86.

W. Zhao, R. Chellappa, and A. Krishnaswamy, Discriminant analysis of principal components for face recognition, Proc Third IEEE Int Conf Automatic Face and Gesture Recogn, Nara, Japan, 14–16 April, 1998, pp. 336–341.

W. Zhao, R. Chellappa, J. Phillips, and A. Rosenfeld, Face recognition in still and video images: A literature survey, ACM Comput Surv 35 (2003), 399–458.