Statistics in Face Recognition: Analyzing Probability Distributions of PCA, ICA and LDA Performance Results

Kresimir Delac [1], Mislav Grgic [2] and Sonja Grgic [2]
[1] *Croatian Telecom, Savska 32, Zagreb, Croatia, e-mail: kdelac@ieee.org*
[2] *University of Zagreb, FER, Unska 3/XII, Zagreb, Croatia*

Abstract

In this paper we address the issue of evaluating face recognition algorithms using descriptive statistical tools. By using permutation methodology in a Monte Carlo sampling procedure, we investigate recognition rate results probability distributions of some well-known algorithms (namely, PCA, ICA and LDA). With a lot of contradictory literature on comparisons of those algorithms, we believe that this kind of independent study is important and will serve to better understanding how each algorithm works. We show how simplistic approach to comparing these algorithms can be misleading and propose a full statistical methodology to be used in future reports. By reporting detailed descriptive statistical results, this paper is the only available detailed report on PCA, ICA and LDA comparative performance currently available in literature. Our experiments show that the exact choice of images to be in a gallery or in a probe set has great effect on recognition results and this fact will further emphasize the importance of reporting detailed results. We hope that this study will help to advance the state of experiment design in computer vision.

1. Introduction

Face recognition is currently highly researched area of computer vision and pattern recognition [1]. While many algorithms are being developed, they are usually compared to existing ones quite superficially and only simple comparisons are reported. Given the numerous theories and techniques that are applicable to face recognition, it is clear that detailed evaluation and benchmarking of these algorithms is crucial. Effort done by FERET researchers in their evaluations [2] pushed face recognition algorithm comparisons to the next level. A common data set and a common testing protocol was designed and other researchers adopted it in their comparisons. Cumulative Match Score (CMS) curve was introduced as a main tool for comparisons. Recognition rate for different algorithms is plotted as a function of rank (rank actually showing among how many top matches is the correct answer) and curves higher in that plot were considered to be superior to other. Using their setup, researchers started to present their results in this new manner, thus usually giving rank 1 results in a table form or plotting CMS curves. With common FERET database used in comparisons researchers were able to reproduce each other's results and compare them directly.

Another important moment in history of face recognition algorithms comparisons was when some researchers concluded that by testing algorithms with FERET protocol, one will not answer some important questions like:

- *which of the measured differences in algorithm performance were statistically distinguishable, and which essentially a matter of chance?* [3]
- *how much does recognition rate vary when comparing images of individuals taken on different days using the same camera?* [4]
- *does one algorithm perform significantly better than another relative to the variance induced by perturbing gallery and probe images?* [4]
- *are recognition results significant with respect to the probe, gallery or training set size?*

In this paper we would like to continue on a framework posed by Beveridge et al. [3], [4] and evaluate *how much effect (variation) the exact images chosen to be in a gallery or in a probe set have on algorithm performance*. The question of images in the training set being representative of a larger population, though important and well known, will not be addressed here. We will compare well known algorithms: Principal Component Analysis (PCA) [5], Independent Component Analysis - Architecture 1 and 2 (ICA1 and ICA2 in further text) [6], and Linear Discriminant Analysis (LDA) [7] combined with common distance metrics (L1, L2, Mahalanobis - *mah* and Cosine - *cos* distance) in a classical nearest neighbor matching algorithm. With this simple example we will present the idea of full methodology for face recognition algorithm comparisons by using advanced descriptive statistical tools and give a first detailed statistical analysis of these algorithms in a single paper currently available in literature. Performance variability due to variations in gallery as well as in probe sets will be investigated to cover as much practical real-life circumstances as possible. It will be shown that these algorithms, when subject to rigorous permutation

testing and statistical analysis, yield much lower results than the ones reported in papers so far.

The goal of this paper is to provide a statistical basis for drawing conclusions about the relative performance of different algorithms to be able to better explain why algorithms behave as they do. This will improve further understanding of traditional methods as well as provide a basis for testing if novel algorithms are really better.

The rest of this paper is organized as follows: previous work is described in Section 2, methodology that we will use in this comparison is described in Section 3, experimental setup is in Section 4, results of detailed PCA, ICA and LDA comparison is given in Section 5 and Section 6 concludes the paper.

2. Previous work

FERET 1996 & FRVT 2000 [2]. The primary tool developed to test face recognition algorithm performance was the CMS curve. Curves higher in the plot represent algorithms "doing better". This measure is highly questionable because even when curves appear higher when visually inspected it can turn out that there is no *significant statistical difference* in performance when compared to lower ones using some sort of statistical hypothesis testing (as proven in [8] and [9]). In a common FERET protocol, algorithms were evaluated against different categories of images. The categories were broken out by lighting change and the time between the acquisition date of the gallery image and probe image. By listing performance in these categories, a better understanding of the face recognition field in general, as well as the strengths and weaknesses of an individual algorithm was obtained. To address the issue of performance variation, the gallery (originally constructed of 1196 images) was partitioned into six galleries of approximately 200 images, in which an individual was in only one gallery. Those galleries were then tested against two standard probe sets (*fb* and *dupII*) and it was concluded that algorithm performance is dependent on the gallery and probe sets. Actually what was done is changing galleries while keeping the probe set fixed. Average range between maximum and minimum performance results for a given gallery set was roughly 13% in recognition rate for *fb* probe set and 45% for *dupII* set. As a guideline for further research the need for measuring the effect of changing galleries and probe sets and statistical measures that characterize these variations was emphasized.

FRVT 2002 [10]. In addition to computing standard performance statistics, new statistical methods were developed to estimate variation in performance over multiple galleries and to explore the effect of covariates on performance. How verification performance varies under two conditions was examined. The first is how performance varies with different galleries. This models the performance of a system that might be installed at different locations. The second is how performance varies for different classes of probes. These results were, unfortunately, reported only for algorithms working in verification mode. Performance was quantified by Receiver Operating Curves (ROC) and uncertainty in results was plotted as error ellipses on those plots. Twelve disjoint galleries and corresponding probe sets were generated from a larger population (similar to FRVT 1996). It was concluded that in face recognition and biometrics, it is an active area of research to develop techniques to measure the uncertainty of an estimator. Both this and FRVT 1996 & FRVT 2000 guideline for further research inspired us to try to contribute to this matter.

FRVT tests never used advanced statistical tools in their performance analyses and this was correctly noticed by Beveridge et al. in [3]. In [3] and [4] distributions and confidence intervals were introduced as an effort to improve current approaches to evaluating face recognition systems. Hypothesis testing was then introduced in [8] as a final step towards including advanced statistics in algorithm comparisons. This paper will not address the issues dealing with hypothesis testing and an interested reader is referred to [9]. In [11] Micheals et al. took a different approach than the one used in [3] and [4]. Balanced sampling was introduced but it was proven by Beveridge [4] that there is no significant difference between balanced and unbalanced approach. This is why we decided to use the unbalanced resampling in our experiments.

3. Statistics to be used in our comparisons

We will use descriptive statistics for comparing mentioned algorithms. After permuting gallery and probe sets in two disjoint experiments (as described in Section 4), rank 1 recognition rate percentage results will be given in a table form consisting of the following advanced statistical data: mean value, standard error, median, mode, standard deviation, sample variance, skewness, kurtosis, range, minimum and maximum value. These data will be given for every algorithm implementation tested. Rank 1 recognition rates distributions for every algorithm will be given in a histogram plot. Since this turned out to be representative enough for larger population (and higher ranks as well) we will not plot all CMS curves in this paper. Relative rankings of algorithms and all their implementations are the same for all ranks (the CMS curves, when plotted, never cross). This is why we will give CMS only for the best implementation of each algorithm.

Our work will differ from FERET 1996 & FRVT 2000 in the fact that we will permute both gallery and probe sets and that our individual sets will not be disjoint (for details see Section 4). We will analyze algorithms in identification mode and this fact will differ our work from the work done in FRVT 2002 (no

descriptive statistics is used in FRVT 2002 neither). Beveridge et al. [4] investigated only PCA and LDA and in not as much detail (regarding descriptive statistics) as we will present here while we expand their work with ICA implementations and give results in a single paper. We will record results to rank 80 as opposed to their rank 10. Since we will not address the issue of when the images were taken we are able to permute images in completely random manner and consider each image equal and exchangeable with any other image of the same individual. This will ensure robustness of our evaluation and our results will represent overall performance.

4. Experimental setup

We used two sets of images from FERET database, which we will name SET3 and SET4. SET3 consists of images of 225 persons for which there are *exactly* 3 images per person in the database and SET4 consists of images of 256 persons for which there are *exactly* 4 images per person. There is no overlap between these two sets. To test algorithms in most unfavorable conditions (i.e. to disable the system to be "tuned" to a specific set of images) we decided to train all algorithms with SET3, leaving SET4 intact to serve as a query set in our virtual experiments. It is worth mentioning that by having 3 images per person, SET3 is probably not ideal for LDA and that is why our LDA results are somewhat inferior to those that would be obtained by using, for example, 15 or more images per person. It also stays unclear if having so many individuals in the training process favors LDA (as mentioned in [3]). All images used in this experiment were first preprocessed using standard steps (spatial transformations, cropping, histogram adjusting to the range of values from 0 to 255). After this, all images were resized to be the size of 60×50 pixels. A standard top 40% of principal components were retained from PCA and this (270-dimensional) subspace was the input for both ICA and LDA algorithms. PCA, ICA1, ICA2 and LDA will not be described here since the idea of this paper is to present methodology for algorithm comparisons and compare those algorithms using advanced statistical tools. We performed two independent experiments; one having fixed gallery set and permuted probe sets and the other having fixed probe set and permuted gallery sets. We achieved this by taking (by random) one image of each person in SET4 and putting it in a gallery. The probe sets were then derived from the remaining 3 images of each person by randomly choosing one image and repeating this process a 100 times. This way we produced a 100 different probe sets. Similar process was used for second experiment with one distinction being that the probe set was fixed and gallery images permuted.

Since we posed a different question then the one presented in [4], we were able to use a larger number of classes in our experiments. Our experiment is more general and we do not need to preserve the multiple day separation information. However, we only tested 100 permutations since this setup turned out to produce results indistinguishable from the experiment we conducted using 1000 permutations for rank 1 results. It is important to mention that the problem of balanced or unbalanced sampling addressed in [4] also does not exist in our methodology. Using the fact that once the systems are trained the distance between any pair of images is constant, we were able to perform virtual experiments. We projected all images from SET4 onto each subspace and calculated distance matrices (distance matrix being the size $n \times n$, where n is the number of images in SET4, and containing the distances between all images). All experiments were then performed on those matrices by changing the list items in probe and gallery sets and without running the algorithms again. For each of the 100 trials in each experiment we recorded recognition results for ranks 1 to 80 and statistically analyzed them.

5. Results for PCA, ICA and LDA

Results of detailed descriptive statistical evaluation of PCA, ICA and LDA can be seen in Table 1 and 2. At first glance it is obvious that all mean rank 1 results are somewhere between 60% and 70%. This is much lower than results reported so far and we believe that this is due to robustness of our methodology (images taken under different lighting and images taken on different days were all mixed and considered exchangeable). Since we considered gallery and probe images of each person exchangeable, results seen in Figure 1 and 2 are equivalent (regarding the distribution) to those that would have been obtained in any hypothetical experiment using different probe and gallery images for these people, i.e. results are representative for a larger population. With our Monte Carlo-like sampling technique we approximated the probability distribution of recognition rates. Recognition rates are histogramed in 16 bins, equally distributed between minimum and maximum recognition rate percentage for a given algorithm.

Looking at Figure 1 and 2 it is obvious that overall results for each algorithm implementation are very similar in both experiments (fixed gallery and fixed probe set). For PCA, we can conclude that L1 metric gives best results (in Figure 1, the results for PCA+L1 are clustered around a bit higher recognition rates), L2 and *cos* are almost indistinguishable and *mah* is the worst choice. ICA1 works almost the same in all implementations but *mah*, and the same can be concluded for LDA also. ICA2 with *cos* metric outperforms by far all other algorithms.

Table 1. Detailed statistical results for fixed gallery and 100 permuted probe sets (all values given in Table are for recognition rate percentage but the percentage sign is omitted)

	PCA				ICA1				ICA2				LDA			
	L1	L2	cos	mah	L1	L2	cos	mah	L1	L2	cos	mah	L1	L2	cos	mah
Mean	69.41	67.02	67.07	55.99	67.61	66.83	66.18	55.99	66.82	66.57	74.50	55.99	67.53	66.61	66.88	60.05
Standard Error	0.22	0.23	0.22	0.22	0.24	0.23	0.23	0.22	0.20	0.20	0.20	0.22	0.21	0.22	0.22	0.21
Median	69.53	67.19	67.19	55.86	67.38	66.80	66.21	55.86	66.80	66.80	74.22	55.86	67.58	66.80	67.19	59.96
Mode	69.92	65.63	67.58	55.86	65.23	68.36	65.63	55.86	66.80	64.84	73.83	55.86	66.80	67.97	66.41	61.72
Standard Deviation	2.16	2.26	2.19	2.18	2.35	2.34	2.30	2.18	1.99	1.97	2.00	2.18	2.15	2.25	2.21	2.09
Sample Variance	4.65	5.11	4.81	4.75	5.52	5.46	5.28	4.75	3.96	3.88	4.01	4.75	4.61	5.04	4.89	4.37
Kurtosis	0.61	0.47	0.07	0.35	-0.47	-0.07	0.03	0.35	-0.38	-0.18	0.01	0.35	0.25	0.22	-0.11	-0.32
Skewness	0.08	-0.17	0.13	0.29	0.10	0.06	-0.10	0.29	-0.01	0.30	0.48	0.29	-0.17	-0.11	0.01	0.19
Range	13.28	13.28	11.33	11.33	11.33	12.50	12.50	11.33	8.59	9.77	9.38	11.33	12.50	12.50	11.33	10.16
Minimum	62.50	59.38	61.33	50.78	62.11	60.16	58.98	50.78	62.11	62.50	69.92	50.78	60.55	59.38	60.55	55.86
Maximum	75.78	72.66	72.66	62.11	73.44	72.66	71.48	62.11	70.70	72.27	79.30	62.11	73.05	71.88	71.88	66.02

Table 2. Detailed statistical results for fixed probe and 100 permuted gallery sets (all values given in Table are for recognition rate percentage but the percentage sign is omitted)

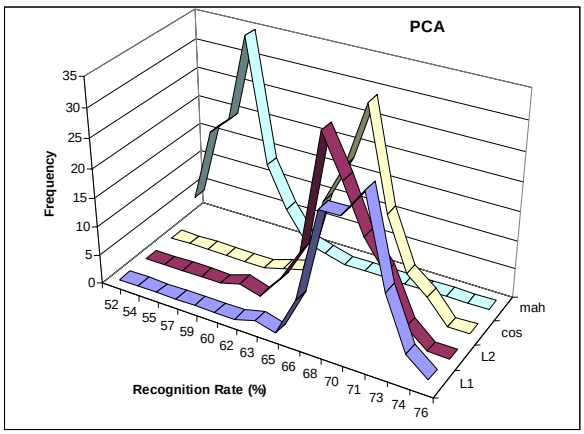
	PCA				ICA1				ICA2				LDA			
	L1	L2	cos	mah	L1	L2	cos	mah	L1	L2	cos	mah	L1	L2	cos	mah
Mean	69.04	66.80	67.63	51.82	65.54	66.16	66.81	51.82	60.40	60.67	74.16	51.82	66.49	66.75	67.38	56.09
Standard Error	0.24	0.26	0.24	0.37	0.24	0.25	0.25	0.37	0.29	0.28	0.21	0.37	0.26	0.26	0.23	0.33
Median	69.14	67.19	67.77	51.95	65.43	66.21	67.19	51.95	60.55	60.74	74.22	51.95	66.41	66.99	67.38	56.64
Mode	67.19	67.58	69.53	51.17	63.67	63.67	67.19	51.17	60.16	58.59	74.22	51.17	64.84	68.75	66.41	58.59
Standard Deviation	2.37	2.60	2.43	3.72	2.43	2.54	2.45	3.72	2.91	2.77	2.12	3.72	2.56	2.55	2.31	3.29
Sample Variance	5.59	6.77	5.88	13.81	5.92	6.47	6.01	13.81	8.46	7.68	4.50	13.81	6.54	6.52	5.34	10.80
Kurtosis	-0.06	-0.49	-0.49	-0.18	-0.09	-0.66	-0.79	-0.18	-0.51	-0.39	-0.09	-0.18	-0.14	-0.61	-0.65	-0.10
Skewness	-0.07	-0.02	0.05	-0.41	0.08	0.11	-0.01	-0.41	-0.28	-0.21	0.28	-0.41	0.15	0.04	0.21	-0.11
Range	12.89	12.11	10.94	17.97	12.89	10.94	10.55	17.97	12.50	12.11	9.77	17.97	12.11	11.33	9.38	17.19
Minimum	62.11	60.94	62.11	40.63	58.59	60.94	61.72	40.63	53.52	53.91	69.92	40.63	60.94	61.33	63.28	47.27
Maximum	75.00	73.05	73.05	58.59	71.48	71.88	72.27	58.59	66.02	66.02	79.69	58.59	73.05	72.66	72.66	64.45

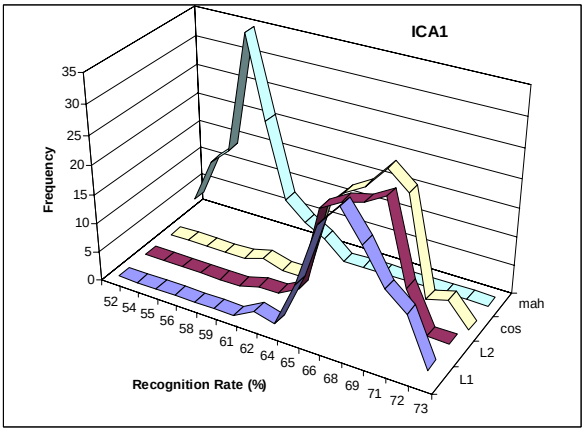
By looking at the figures and kurtosis and skewness in tables a lot can be concluded about the distributions of results for each individual implementation. For example, we can see that ICA2+*cos*, by having kurtosis (Table 1, first experiment) close to zero (0.01), has distribution close to normal Gaussian as far as the "peakness" is concerned, and by having positive and fairly large skew (0.48), has its "right" (or "positive", if we consider mean value to be zero) tail longer than its "left", meaning that the values right of the mean have higher scatter. All this is clearly illustrated in Figure 1 as well.

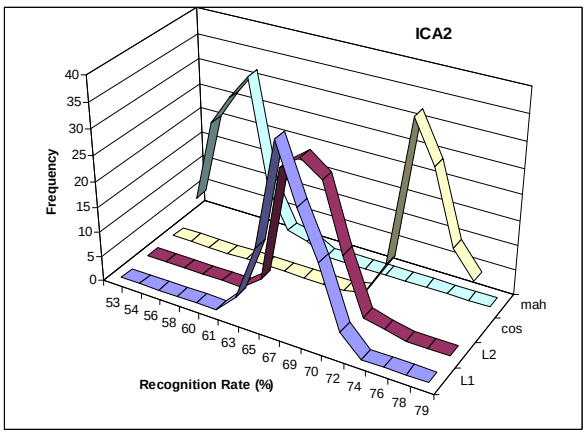
Taking PCA+L1 from Table 1 as an example, we can show how conducting one experiment and with just one gallery and probe set can be misleading. We can see that the minimum recognition rate for this implementation was 62.50% and the maximum 75.78%. Reporting either one of these two values (because you did only one experiment) actually gives no insight into the real underlying distribution of algorithm's results. 62.50% would be unfairly low and 75.78% unrealistically high.

Mean CMS curves for the best implementations of each algorithm are shown in Figure 3 for the first experiment and in Figure 4 for the second. As expected, relative rankings of algorithms are the same for both experiments and the best one seems to be ICA2+*cos*, although PCA and LDA have similar results at higher ranks. Hypothesis testing should be used here to determine when exactly is the difference between ICA2+*cos* and other algorithms statistically significant (please refer to [9]). ICA1 is clearly inferior and this once again shows that ICA1 should be used when recognizing facial actions [6] rather than in classical face recognition setup.

Our results seem to contradict some reported comparisons. Most obvious being that LDA was shown to be superior to others in FERET evaluations and this is certainly not the case here (possible reasons are given in Section 4), as shown also in [12]. With our more strict and robust testing methodology we confirmed Bartlett's et al. [6] conclusions (which they made using a simple FERET-like testing methodology) that ICA2 is the best currently available algorithm.







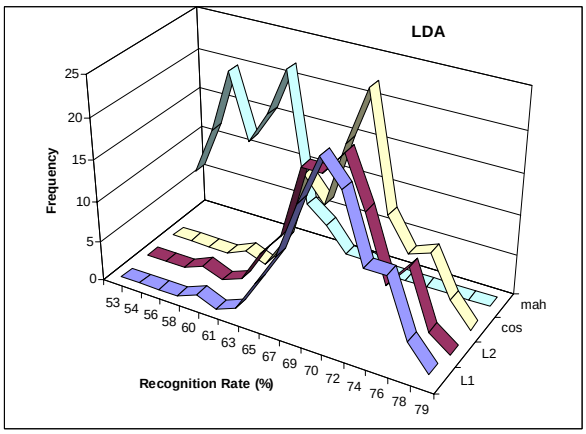
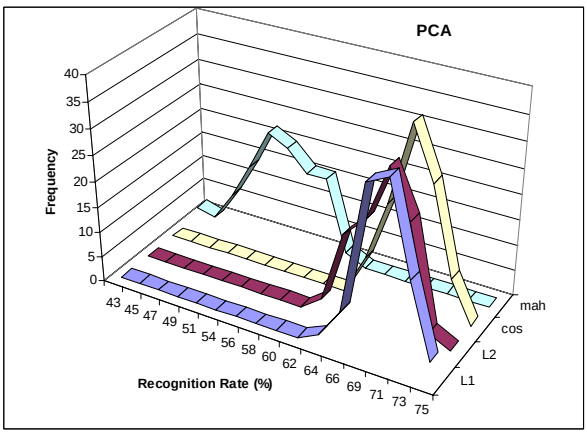
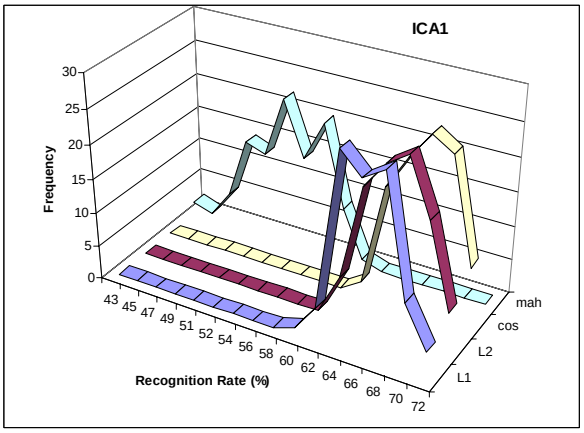
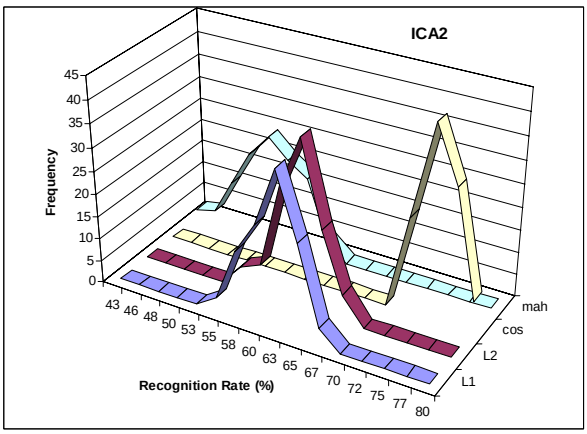


Figure 1. Recognition Rate Distribution
(fixed gallery, permuted probe sets)







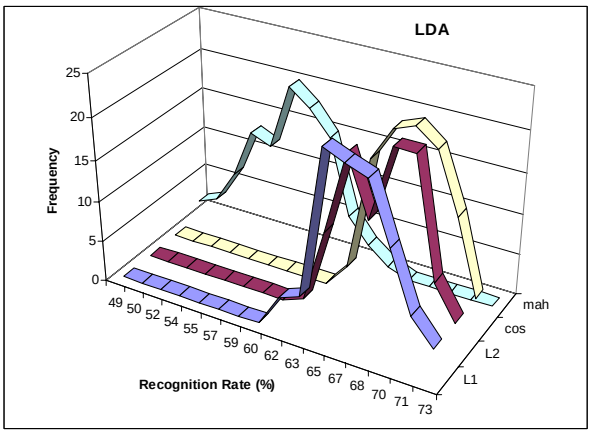


Figure 2. Recognition Rate Distribution
(fixed probe, permuted gallery sets)

6. Conclusions

In this paper we reviewed some major issues concerning statistical evaluation of face recognition algorithms. We presented a complete methodology for algorithm comparisons and illustrated it by evaluating well-known algorithms (PCA, ICA1, ICA2 and LDA) in four possible implementations while working in identification mode. Algorithms were tested on the same set of images by permuting the choice of gallery and probe images in a Monte Carlo study. It was concluded that the exact choice of images to be in a gallery or in a probe set has great effect on recognition results. This is why we believe that reporting detailed statistical analysis when comparing face recognition algorithms is important for other researcher to be able to get an insight into the distribution of performance results. We hope that experiments and results described in this paper sufficiently illustrate the importance of the presented approach and will help researches in their desire to know how algorithms behave when changes are made to gallery and probe sets. We believe that the presented methodology could be used to augment the existing evaluation trends and help to advance the state of experiment design in computer vision.

Acknowledgment

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

References

[1] W. Zhao, R. Chellappa, J. Phillips, A. Rosenfeld, "Face Recognition in Still and Video Images: A Literature Survey", *ACM Computing Surveys*, Vol. 35, Dec. 2003, pp. 399-458

[2] P.J. Phillips, H. Moon, S.A. Rizvi, P.J. Rauss, "The FERET Evaluation Methodology for Face Recognition Algorithms", *IEEE Trans. on PAMI*, Vol. 22, No. 10, October 2000, pp. 1090-1104

[3] 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. of the IEEE Conference on Computer Vision and Pattern Recognition*, Kaui, HI, USA, December 2001, pp. 535- 542

[4] J.R. Beveridge, K. She, B. Draper, and G. Givens. "Parametric and Non-parametric Methods for the Statistical Evaluation of HumanID Algorithms", *IEEE Third Workshop on Empirical Evaluation Methods in Computer Vision*, Kauai, HI, USA, December 2001

[5] M. Turk, A. Pentland, "Eigenfaces for Recognition", *Journal of Cognitive Neuroscience*, Vol. 3, No. 1, 1991, pp. 71-86

[6] M.S. Bartlett, J.R. Movellan, T.J. Sejnowski, "Face Recognition by Independent Component Analysis", *IEEE Trans. on Neural Networks*, Vol. 13, No. 6, November 2002, pp. 1450-1464

[7] P. Belhumeur, J. Hespanha, D. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection", *Proc. of the Fourth European Conference on Computer Vision*, Vol. 1, Cambridge, UK, 1996, pp. 45-58

[8] W. Yambor, B. Draper, R. Beveridge, "Analyzing PCA-Based Face Recognition Algorithms: Eigenvector Selection and Distance Measures", *Empirical Evaluation Methods in Computer Vision*, H. Christensen and J. Phillips, eds., World Scientific Press, 2002

[9] K. Delac, M. Grgic, S. Grgic, "Independent Comparative Study of PCA, ICA, and LDA on the FERET Data Set", Technical Report, University of Zagreb, FER, 2004

[10] P.J. Phillips, P. Grother, R.J. Micheals, D.M. Blackburn, E. Tabassi, M. Bone, "Face Recognition Vendor Test 2002 - Evaluation Report", 2003 http://www.frvt.org/DLs/FRVT_2002_Evaluation_Report.pdf,

[11] R.J. Micheals, T Boult, "Efficient Evaluation of Classification and Recognition Systems", *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 1, 2001, pp. 50-57

[12] A. Martinez, A. Kak, "PCA versus LDA", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 23, No. 2, February 2001, pp. 228-233

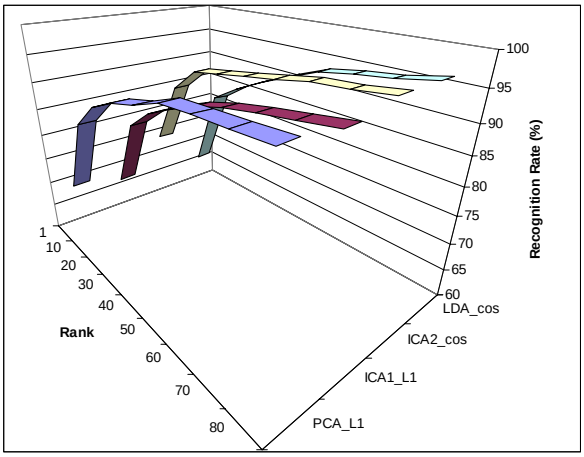


Figure 3. CMS for best implementations (fixed probe, permuted gallery sets)

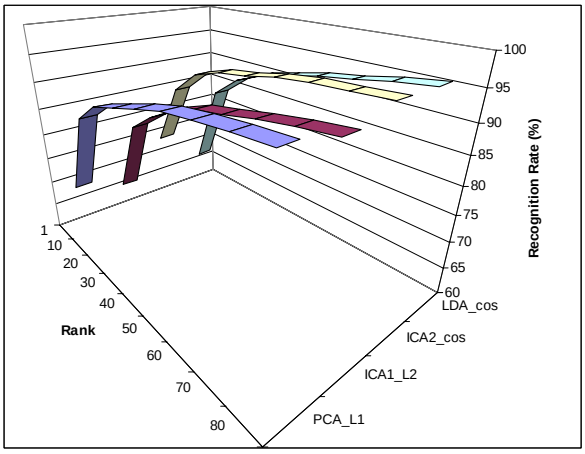


Figure 4. CMS for best implementations (fixed gallery, permuted probe sets)