Generalization Abilities of Appearance-Based Subspace Face Recognition Algorithms

Kresimir Delac *, Mislav Grgic and Sonja Grgic
Department of Wireless Communications, Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia
E-mail: kdelac@ieee.org
* Corresponding author

Abstract: In this paper we present an efficient method to test the generalization abilities of subspace face recognition algorithms. The main motivation for this work is the lack of detailed analysis of this problem in current literature. Generalization ability of face recognition algorithm is the ability to recognize new individuals, which were not part of the training process. To illustrate our idea we used well-known recognition algorithms (PCA, ICA and LDA) and the FERET date set. Our results show that even these well-known algorithms have poor generalization abilities in some implementations.

Keywords: Face Recognition, PCA, ICA, LDA, FERET, Generalization Abilities.

Reference to this paper should be made as follows: Delac, K., Grgic, M. and Grgic, S. (2005) 'Generalization Abilities of Appearance-Based Subspace Face Recognition Algorithms', *12th Int. Workshop on Systems, Signals & Image Processing,* 22-24 September 2005, Chalkida, Greece

Biographical notes: K. Delac received his B.Sc.E.E. degree in 2003 and is currently pursuing a Ph.D. degree at the University of Zagreb, Faculty of Electrical Engineering and Computing. His current research interests are digital image analysis, pattern recognition and biometrics.

M. Grgic is an Assistant Professor at the University of Zagreb, Faculty of Electrical Engineering and Computing. He received his Ph.D. degree in electrical engineering from University of Zagreb, Faculty of Electrical Engineering and Computing, Zagreb, Croatia, in 2000. His current research focuses on image and video compression (DCT, wavelets), image retrieval (MPEG-7), face recognition / biometrics and multimedia communications.

S. Grgic is an Associate Professor at the University of Zagreb, Faculty of Electrical Engineering and Computing. She received her Ph.D. degree in electrical engineering from University of Zagreb, Faculty of Electrical Engineering and Computing, Zagreb, Croatia, in 1996. Her current research focuses on television signal transmission and distribution, picture quality assessment, wavelet image compression, and broadband network architecture for digital television.

1 INTRODUCTION

Face recognition (Zhao et al., 2003), and biometrics in general (Jain et al., 2004), became one of the most important research questions in recent years. Pioneering days are way behind us and at this stage there is a significant need for deeper understanding and thorough analysis of existing algorithms. One property rarely addressed in evaluations is the generalization ability of an algorithm. There are several different meanings of the expression *generalization ability* currently in use. It is sometimes addressed as an ability to maintain a recognition rate when reducing the number of images in the training set. Another meaning would be, when considering the algorithms that use more than one image per class in training, the ability of an algorithm to maintain a recognition rate when the number of images per class used in training is reduced (Navarrete and Ruiz-del-Solar, 2002). For recognition of faces under various pose, generalization is the ability to recognize faces under poses that were not used in training (the same thing can be said for facial expressions as well). In our work we are focusing on the definition of generalization abilities as being the algorithm's ability to recognize images that were not part of the training process. Since we have used well-known subspace face recognition algorithms and FERET (Phillips et al., 2000) database images and nomenclature to illustrate our idea, we can restate the definition as *the ability to recognize individuals that were not part of the training set when computing the subspace*. Being more precise, we investigate the effect that an overlap between images used as a training set (T) and images used in gallery (G) has on recognition rate. The generalization ability of a specific algorithm is better if recognition rate does not depend on overlap between T and G. Our results show that, out of 12 tested algorithms, 4 algorithms show significant differences at rank 1 recognition rates across various overlaps of T and G and thus have poor generalization abilities.

The rest of this paper is organized as follows: Section 2 gives a brief overview of algorithms, Section 3 describes experimental setup, Section 4 reports the results and Section 5 concludes the paper.

2 SUBSPACE FACE RECOGNITION

We used three most popular subspace projection methods currently used in face recognition to illustrate our idea:

PCA. Given an s-dimensional vector representation of each face in a training set of images, Principal Component Analysis (PCA) (Turk and Pentland, 1991) 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 ($t << s$). If the image elements are considered as random variables, the PCA basis vectors are defined as eigenvectors of the scatter matrix.

ICA. Independent Component Analysis (ICA) (Bartlett et al., 2002) 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. 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).

LDA. Linear Discriminant Analysis (LDA) (Belhumeur et al., 1996) 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. 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 are the eigenvectors of ($S_W^{-1} \cdot S_B$).

Metrics. We combine three well-known metrics with the described projection methods: L1, L2 and cosine distance (C in out experiments), thus yielding 12 different algorithms (projection-metric combinations). Nearest neighbour algorithm is used in the matching stage.

3 EXPERIMENTAL SETUP

We used standard FERET database. 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.

To obtain the information we need about recognition rates of projection-metric combinations (algorithms), we made five different test sets for five different percentages of overlap between T and G (0%, 25%, 50%, 75% and 100%). 0% of overlap means that no image from G is part of T (actually, even no class from G is part of T in our case). 100% of overlap means that all images from G were used in T as well. Probe set (P) is always a set of images different from T and G.

FERET dataset consists of 3,816 images of 1,201 classes (different persons). We decided to use three images per class in training since that is the bear minimum for LDA to work properly. Thus, for our needs we have chosen two sets of images from the database: 1) images of those persons for which there are three images/class (SET_3) and 2) images of those persons for which there are four images/class (SET_4). 130 classes were randomly chosen in the SET_3 and another 130 different classes in the SET_4. Consequently, there were 3 × 130 = 390 images in SET_3 and 4 × 130 = 520 images in SET_4. We needed those two separate sets to achieve different overlaps. For example, the 25% overlap was achieved by taking 98 classes (3 × 98 = 294 images) from SET_3 and 32 classes (3 × 32 = 96 images) from SET_4. Since we use three images per class for training, we always have one image per class from SET_4 never used in training, which is used in the probe set. After training, we take one image per class of those 3 × 32 images from SET_4 that were used in T and one image from the rest of 98 classes from SET_4 and thus yield a gallery of 130 images (one image per class). This gives us the overlap between T and G of 25% (32/130 ≈ 0.25). From the explanation above it is obvious that, apart from the images that overlap in T and G, the rest of the training set consists of persons that are not used in the recognition stage. This way we actually test if an algorithm is tuned to a specific gallery or a specific set of people. From those 3 × 130 = 390 images used in training, PCA derived 389 meaningful eigenvectors. We decided to keep the top 40% of those, corresponding to largest eigenvalues. This way a new 160 dimensional subspace was derived (160/389 ≈ 0.41). PCA based algorithms used this space for recognition. This space was also the input for ICA, which yielded a 160 dimensional space as well (both for ICA1 and ICA2). LDA, however, yielded only 129 dimensional space since it can produce a maximum of c – 1 basis vectors (c being the number of classes). All of those were kept to stay as close as possible to the dimensionality of PCA and ICA.

4 RESULTS

Results of our experiments are shown in Table 1 and Figure 1. Table 1 presents recognition rate results at rank 1 for 12 tested algorithms at a given percentage of T and G overlap (0% – 100%). These results are also presented in Figure 1. Our hypothesis is: closer that graph in Figure 1 is to a horizontal line, the better the generalization abilities of a specific algorithm are. We can see that, although the difference in algorithm's performance is not extreme, it is noticeable. Before we can draw any further conclusions we need to determine if the algorithm's performance is really significantly different for different overlaps, because we can not make any strong conclusions from the visual inspection of the results presented in Figure 1. So, we decided to use hypothesis testing. Two testing techniques were used; a classic statistical *z*-test and a newly introduced *McNemar's test* (Yambor et al., 2002; Yang et al., 2005). The results are shown in Table 2. We decided to test two extreme results for a given algorithm (minimum and maximum number of images correctly recognized – *Score* in Table 2) and see if the difference between algorithm's performance at overlaps for which those results were achieved is significant (the *S* columns in Table 2). *z*-test will give us the answer to the question: *is the difference significant considering the probe set as a whole*? To perform this test we pose our hypotheses as follows: *H1*) Algorithm correctly recognizes images more often at overlap X then at overlap Y, and *H0*) There is no difference in how it performs at different overlaps. To be able to claim that *H1* is true we need to establish that the probability of *H0* (*p*(*H0*)) is very small. Variable z is calculated and *p*(*H0*) is determined. It can be seen that, given the standard 0.05 cutoff, the only significant difference in performance across various overlaps is recorded for ICA2+L1. This is also obvious from the plot in Figure 1. However, when we took a closer look at Figure 1 we noticed that the difference in performance could (or should) be significant for some other algorithms as well. Thus, we decided to use a more discriminating McNemar's test. McNemar's test is a null hypothesis statistical test based on a Bernoulli model and it will give us more precise results because it takes full advantage of our experimental protocol (similar to Yambor et al., 2002). We redefined our hypotheses as: *H1*) when algorithm's performances at overlap X and overlap Y differ on a particular image, algorithm is more likely to correctly recognize it at overlap X, and *H0*) when algorithm's performances differ on a particular image at overlap X and overlap Y, algorithm is equally likely to correctly recognize it for both overlaps. For details on both tests please refer to (Yambor et al., 2002). *p*(*H0*) is again calculated and now we can see that four algorithms actually perform significantly different (the same 0.05 cutoff is used).

We can now safely conclude that ICA1+L2, ICA1+C, ICA2+L1 and LDA+L2 have worse generalization abilities than other tested algorithms since they perform significantly different when the percentage of overlap between T and G changes. If we now take a closer look at the values of *p(H0)* for the McNemar's test, we can see that *p(H0)* of four more algorithms (PCA+L2, ICA1+L1, LDA+L1 and LDA+C) is close to rejection since the values of *p(H0)* are of the same magnitude as the 0.05 cutoff. Moreover, algorithms for which the *p(H0)* value is large have very good generalization abilities, and they are: PCA+L1, PCA+C, ICA2+L2 and ICA2+C.

Table 1. *Recognition rates at rank 1*

	Percentage of T and G Overlap				
Alg.	*0%*	*25%*	*50%*	*75%*	*100%*
PCA+L1	67.6%	69.2%	70.0%	67.6%	68.4%
PCA+L2	59.2%	63.8%	65.3%	66.1%	66.1%
PCA+C	61.5%	65.3%	66.9%	66.9%	66.9%
ICA1+L1	58.4%	65.3%	65.3%	66.1%	66.9%
ICA1+L2	59.2%	64.6%	65.3%	69.2%	67.6%
ICA1+C	60.0%	65.3%	65.3%	68.4%	66.1%
ICA2+L1	63.8%	56.9%	53.0%	49.2%	46.1%
ICA2+L2	61.5%	67.6%	66.9%	66.9%	63.8%
ICA2+C	73.8%	80.0%	77.6%	77.6%	80.0%
LDA+L1	59.2%	62.3%	65.3%	63.8%	66.9%
LDA+L2	58.4%	65.3%	66.1%	66.9%	67.6%
LDA+C	61.5%	65.3%	66.9%	66.9%	68.4%

Table 2. *Results of hypotheses testing*

	Score		***z-test***			***McNemar***	
Alg.	*min*	*max*	*z*	*p(H0)*	*S*	*p(H0)*	*S*
PCA+L1	88	91	0.402	0.368	N	0.3679	N
PCA+L2	77	86	1.154	0.205	N	0.0998	N
PCA+C	80	87	0.906	0.265	N	0.1553	N
ICA1+L1	76	87	1.411	0.148	N	0.0631	N
ICA1+L2	77	90	1.682	0.097	N	0.0235	**Y**
ICA1+C	78	89	1.423	0.145	N	0.0401	**Y**
ICA2+L1	60	83	2.867	0.007	**Y**	0.0005	**Y**
ICA2+L2	80	88	1.038	0.233	N	0.1465	N
ICA2+C	96	104	1.178	0.199	N	0.1147	N
LDA+L1	77	87	1.285	0.175	N	0.0717	N
LDA+L2	76	88	1.542	0.121	N	0.0365	**Y**
LDA+C	80	89	1.170	0.201	N	0.0939	N

5 CONCLUSION AND FURTHER WORK

We have presented an efficient approach to evaluating generalization abilities of subspace face recognition algorithms. To illustrate it we used well-known algorithms (PCA, ICA and LDA) and FERET data set. t was shown that, when submitted to this kind of testing, 4 out of 12 tested algorithms gave significantly different recognition results at rank 1 for different T and G overlap percentages. They are namely: ICA1+L2, ICA1+C, ICA2+L1 and LDA+L2. Algorithms PCA+L1, PCA+C, ICA2+L2 and ICA2+C have shown very good generalization abilities, even when subject to this kind of rigorous testing. Visual inspection of the proposed graph is in some cases enough but we strongly recommend the use of statistical hypothesis tests as well. It was shown that classic statistical *z*-test confirms conclusions based on the visual inspection of the graphs, but a newly introduced *McNemar's test* is more discriminating and produces more precise results.

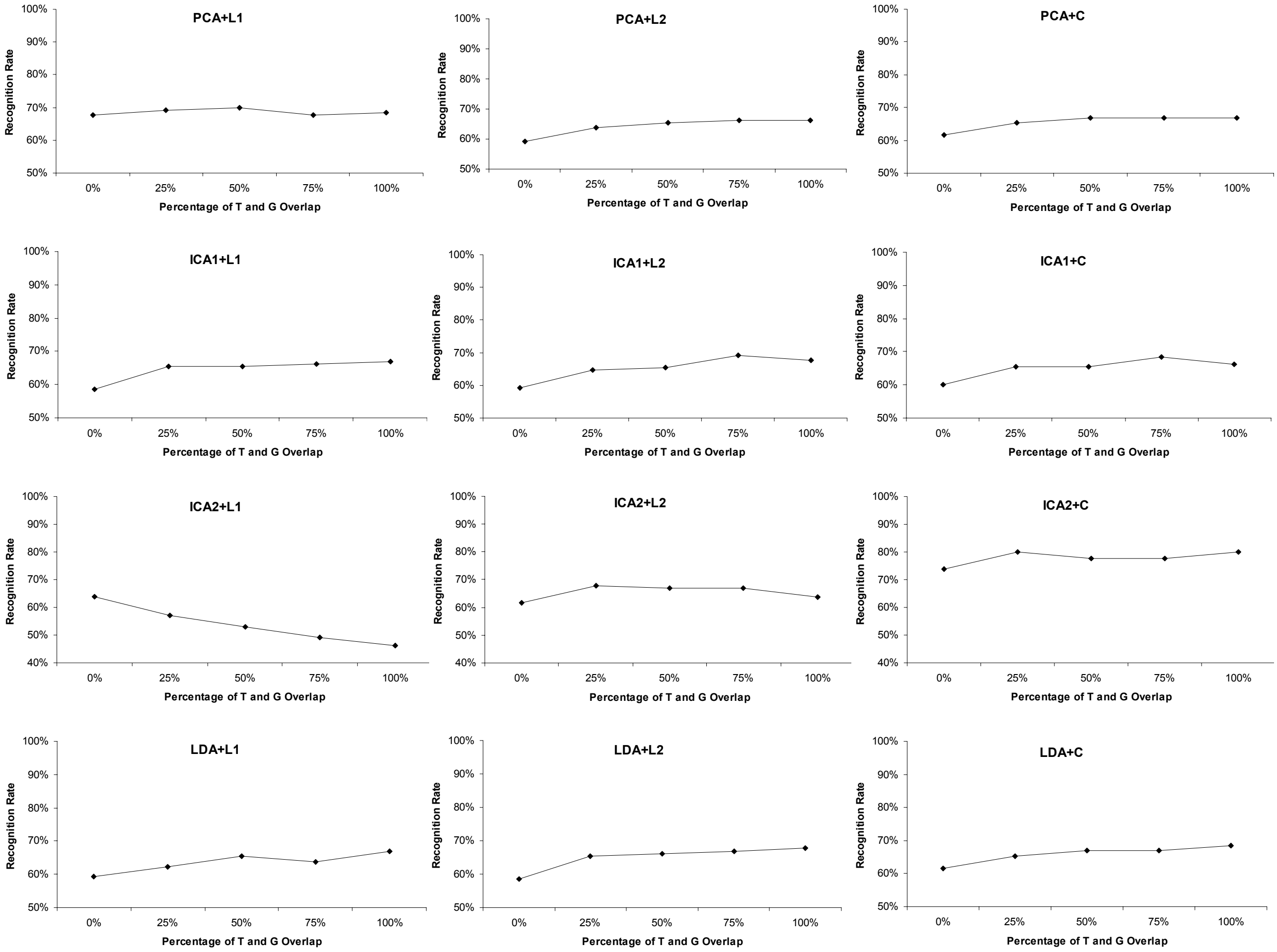

Figure 1. *Plots of rank 1 recognition rates for a specific algorithm at a given percentage of T and G overlap*

We would like to encourage researchers to use similar tests of generalization abilities for their algorithms, when reporting results.

As our further work we would like to use hypothesis testing across all ranks (whole CMS curve) to provide a more detailed analysis. Also, we presume that more precise results would be obtained by permuting a larger number of gallery and probe images once a subspace is defined (after the training step) and stronger conclusions could be drawn.

ACKNOWLEDGMENTS

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

REFERENCES

Bartlett, M.S., Movellan, J.R., Sejnowski, T.J. (2002) 'Face Recognition by Independent Component Analysis', IEEE Transactions on Neural Networks, Vol. 13, No. 6, pp. 1450-1464, November.

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

Jain, A.K., Ross, A., Prabhakar, S. (2004) 'An Introduction to Biometric Recognition', *IEEE Transactions on CSVT*, Vol. 14, No. 1, pp 4-19, January.

Navarrete, P., Ruiz-del-Solar, J. (2002) 'Analysis and Comparison of Eigenspace-Based Face Recognition Approaches', *International Journal of Pattern Recognition and Artificial Intelligence*, Vol. 16, No. 7, pp. 817-830, November.

Phillips, P.J., Moon, H., Rizvi, S.A., Rauss, P.J. (2000) 'The FERET Evaluation Methodology for Face Recognition Algorithms', *IEEE Transactions on PAMI*, Vol. 22, No. 10, pp. 1090-1104, October.

Turk, M., Pentland, A. (1991) 'Eigenfaces for Recognition', *Journal of Cognitive Neuroscience*, Vol. 3, No. 1, pp. 71-86.

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

Yang, J., Frangi, A.F., Yang, J., Zhang, D., Jin, Z. (2005) 'KPCA Plus LDA: A Complete Kernel Fisher Discriminant Framework for Feature Extraction and Recognition', *IEEE Transactions on PAMI*, Vol. 27, No. 2, pp. 230-244, February.

Zhao, W., Chellappa, R., Phillips, J., Rosenfeld, A. (2003) 'Face Recognition in Still and Video Images: A Literature Survey', *ACM Computing Surveys*, Vol. 35, pp. 399-458, December.