Presented at: BMCV*02, Tübingen, FRG, 22-24 November 2002.
Published in: Biologically Motivated Computer Vision, eds.
H.H. Bülthoff, S.-W. Lee, T.A. Poggio and C. Wallraven,
LNCS 2525, pages 491-501, 2002, Springer Verlag, Heidelberg.

Gender Classification of Human Faces

Arnulf B. A. Graf and Felix A. Wichmann

Max Planck Institute for Biological Cybernetics
Spemannstrasse 38, 72076 Tübingen, Germany
{arnulf.graf, felix.wichmann}@tuebingen.mpg.de

Abstract. This paper addresses the issue of combining pre-processing methods—dimensionality reduction using Principal Component Analysis (PCA) and Locally Linear Embedding (LLE)—with Support Vector Machine (SVM) classification for a behaviorally important task in humans: gender classification. A processed version of the MPI head database is used as stimulus set. First, summary statistics of the head database are studied. Subsequently the optimal parameters for LLE and the SVM are sought heuristically. These values are then used to compare the original face database with its processed counterpart and to assess the behavior of a SVM with respect to changes in illumination and perspective of the face images. Overall, PCA was superior in classification performance and allowed linear separability.
Keywords. Dimensionality reduction, PCA, LLE, gender classification, SVM

1 Introduction

Gender classification is arguably one of the more important visual tasks for an extremely social animal like us humans—many social interactions critically depend on the correct gender perception of the parties involved. Arguably, visual information from human faces provides one of the more important sources of information for gender classification. Not surprisingly, thus, that a very large number of psychophysical studies has investigated gender classification from face perception in humans [1]. The aim of this study is to explore gender classification using learning algorithms. Previous work in machine learning focused on different types of classifiers for gender classification—e.g. SVM versus Radial Basis Function Classifiers or Nearest-Neighbor Classifiers—using only low resolution "thumbnail" images as inputs [2]. Here we investigate the influence of two popular dimensionality reduction methods on SVM classification performance using high-resolution images. Ultimately, the success and failure of certain pre-processors and classification algorithms might inform the cognitive science community about which operators may or may not be plausible candidates for those used by humans.

In sec. 2 the MPI human head image database is presented together with the "clean up" processing required to obtain what we refer to as the *processed database*. The dimensionality of its elements is reduced in sec. 3 using PCA and LLE and we look at a number of common summary statistics to identify outliers

and/or see how the choice of PCA versus LLE influences the homogeneity of the reduced "face space". In sec. 4 gender classification of the processed face database in its PCA and LLE representations is studied. The optimal parameters of the SVM (trade-off parameter and kernel function) and LLE (number of nearest neighbors) are determined heuristically by a parameter search. Furthermore, these parameters are used to compare the original to the processed database and to study the dependency of classification on illumination and perspective of the faces.

2 Original and Processed MPI Head Database

The original MPI human head image database as developed and described in [3] is composed of 100 male and 100 female three-dimensional heads. From these, 256x256 color images were extracted at seven different viewing angles (0, ± 9, ± 18 and $\pm 45°$) and three different illumination conditions (frontal, $\Theta = 0, \Phi = 0$; light from above and off center, $\Theta = 65, \Phi = 40$; light from underneath and off center, $\Theta = -70, \Phi = 35$; Θ is the elevation and Φ the azimuth in degrees). The following inhomogeneities in shape and texture can then be observed: on average the male faces are darker and larger than the females and the faces are not centered, with female faces, on average, slightly more offset to the left. In the processing of the database these cues are eliminated since they may be exploitable by an artificial classifier but are, for humans in a real environment, neither reliable nor scientifically interesting cues to gender: we do not normally have a bias to see people as female in the distance (small size), and neither do we have a tendency to see people in the shade (low luminance) as males. Thus the MPI head database was modified in the following way. First, we equalized the intensity of each face to the global mean intensity over all faces. Second, all faces were re-scaled to the mean face size. Finally all faces were centered in the image by aligning the center of mass to the center of the image. The set of faces obtained following the above scheme is referred to below as *processed* head database. Figure 1 shows 4 female and male exemplars of the original and the processed database for comparison. The above processing should be considered as a first step using *any* face database prior to machine classification or psychophysical investigation of gender classification.

3 Pre-processing Using PCA and LLE

Perhaps the first question to arise in machine categorization is the choice of data representation. Images, of faces or natural scenes, contain highly redundant information so a pixel-by-pixel representation appears not suitable. Thus adequate pre-processing in the context of gender classification of faces implies dimensionality reduction. First (truncated) PCA is considered as a benchmark because of its simplicity and wide domain of application [4]. Perhaps more importantly, PCA decomposition, with the eigenvectors with non-zero eigenvalues referred to as *eigenfaces*, has become a strong candidate as a *psychological*

Fig. 1. Comparison between heads from the original database (1st and 3rd columns) and heads from the processed database (2nd and 4th columns).

model of how humans process faces [5–7]. Second its nonlinear neighborhood-preserving extension, LLE, is considered [8, 9]. The latter may be viewed as more biologically-plausible than PCA since it is invariant to rotations, re-scalings and translations: desirable properties for object representation in any biological or biologically-motivated vision system. Here we consider the nearest-neighbor version of LLE since the manifold underlying the face representation cannot be expected to be "smooth" or having a homogeneous sample density. Thus the construction of a local embedding from a fixed number of nearest neighbors appears more appropriate than from a fixed subspace, e.g all neighbors within a hypersphere of fixed radius. We limit the dimensionality of the reduced face space into which PCA or LLE are projecting to 128. In the case of LLE, we consider the 15 nearest neighbors out of a possible maximum of 99, this number being optimal for classification purpose as suggested by the experiments in the next section.

Looking at each of the 200 faces on-screen we find that in "psychological face space" (i.e our perception) no single face appears to be particularly "odd", i.e. the face database seems not to contain outliers. To explore the topography of the PCA- and LLE-induced face spaces the clustering of the elements of the face space is studied by examining the first four moments (mean, variance, skewness and kurtosis) of the distribution of distances between faces. In addition, by iteratively removing the faces corresponding to the tails of the distribution, i.e. the largest outliers, we see how homogeneous the distribution of faces in the

respective face spaces is: large changes in the moments after removal of a small number of exemplars may indicate a sub-optimal pre-processing. In total we removed up to 15 faces for each gender.

The individual contributions—the faces—to the four moments of the processed database are shown in figure 2 for the whole database and with 5, respectively 15, outlying faces removed from it. From figure 2 it can be seen that

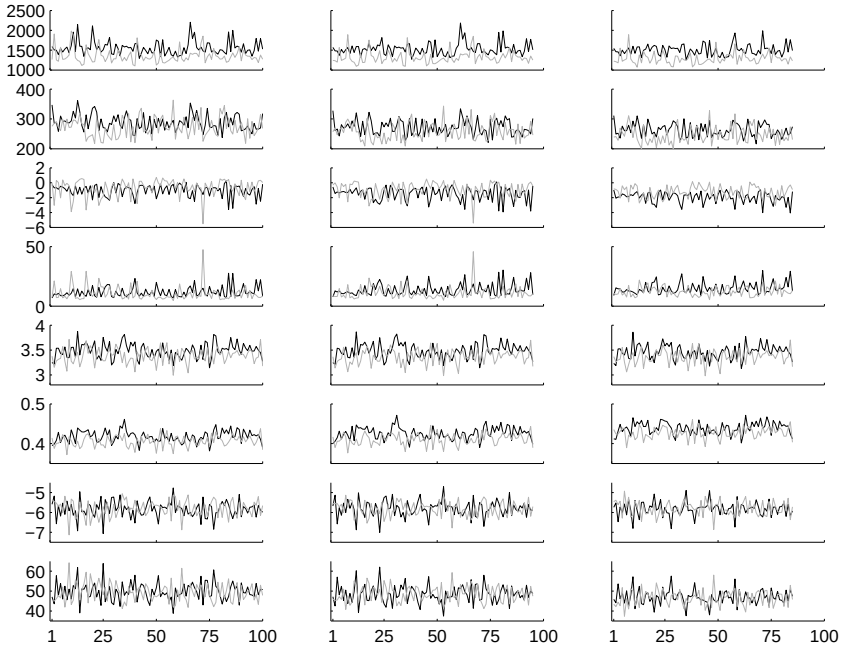

Fig. 2. Comparison of the first four moments (mean, variance, skewness and kurtosis) for each gender based upon PCA (4 first rows) and LLE (4 last rows) for the processed database with (respectively, from left to right) 0, 5, or 15 faces removed. The dark lines correspond to the males and the lighter ones to the females.

for PCA one would have to remove 15% of the elements of the original database in order to eliminate the obvious peaks corresponding to outliers such as the one for skewness around female 70. For LLE, on the other hand, the statistics have clearly fewer peaks, even without removal of outliers, implying a better clustering of the data. This may be explained by recalling that LLE is based upon reconstruction of the data preserving local neighborhoods, and thus also the clusters which may be present in the database. If this analysis is correct, clustering algorithms such as one-class SVMs [10] should show superior clustering ability for LLE data representation than for PCA representation.

4 Classification Using SVMs

The purpose of this section is the study of gender classification in the reduced face space given by PCA or LLE using Support Vector Machines (SVMs, see [11]). The performance of SVMs is assessed though their classification error and the number of Support Vectors (SVs). The kernel functions are normalized and the offset of the optimal separating hyperplane is modified as introduced in [12]. The performance of the SVM is assessed using cross-validation experiments consisting of 100 repeats, each one using 60 random training and 40 random testing patterns for each gender. This 60/40% training/testing subdivision of the dataset was suggested by the study of the standard deviation of the classification error in a preliminary set of experiments.

4.1 Determination of Optimal Parameters

We are confronted with a three-parameter optimization problem: the trade-off parameter c of the SVM, its kernel function and the number of nearest neighbors of LLE. For reasons of computational feasibility, we shall proceed heuristically in the determination of these parameters using the processed database, all values being averaged across the 3 illumination conditions and the 7 perspectives. The first parameter, c, is determined separately for PCA and LLE as shown in figure 3 for a linear and a polynomial kernel of degree 2 respectively. These kernel functions were shown to be optimal during pre-run experiments. In the case of

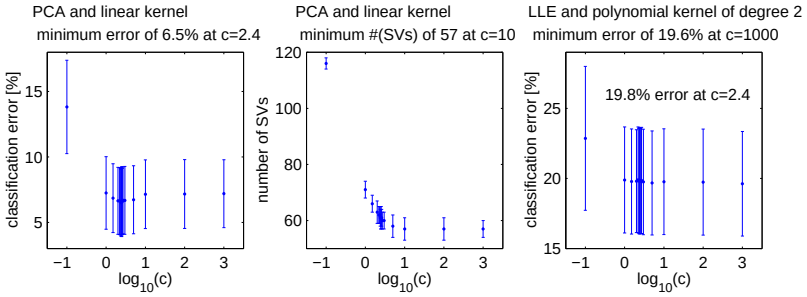

Fig. 3. Mean classification error and number of SVs as function of c using PCA with a linear kernel and LLE with a polynomial kernel of degree 2. In the last case, the number of SVs is found to be constant at a value of 120.

PCA, the value of c obtained for a minimum classification error differs from the one corresponding to a minimum number of SVs, but both values of c are at least of the same order of magnitude. Since in the context of classification a minimum classification error is more relevant than a reduced number of SVs[1], we shall

[1] A reduced number of SVs may be of higher importance than the actual classification error in the context of minimal data representation or data compression.

consider $c_{opt} = 2.4$ as the optimal value of c for PCA in combination with a linear kernel. When doing the same for LLE in combination with a polynomial kernel of degree 2 the classification error curve does not exhibit a global minimum and the number of SVs is constant. We can thus choose c_{opt} as obtained for PCA to be the optimal value of c also in this case. Both classification error curves as function of c exhibits a flat behavior for $1 \leq c \leq 1000$. In this range, the value of c is not of practical importance. This fact combined with the generalization ability of SVMs allows us to extrapolate that the value obtained here for c_{opt} may nearly be also optimal for other kernel functions. However this cannot be guaranteed and this is the price to pay when proceeding heuristically in the three-parameter optimization since a full exploration of these parameters is, alas, computationally prohibitive. The determination of the optimal kernel function of the SVM for PCA and LLE is done by performing classification experiments at c_{opt} as shown in figure 4. From this figure we see that the best performance for PCA comes

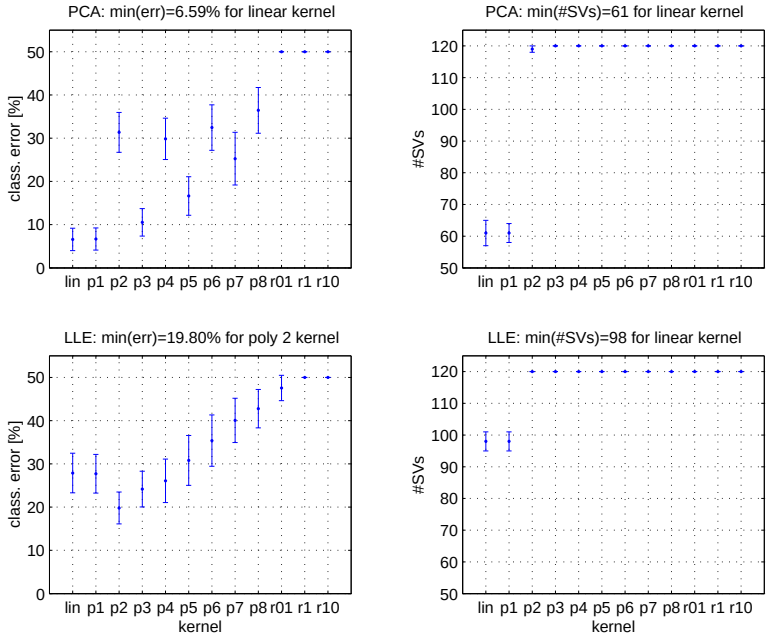

Fig. 4. Classification performance of PCA and LLE with 15 nearest neighbors as function of the kernels: *lin* 1 corresponding to a linear kernel $K(\boldsymbol{x}, \boldsymbol{y}) = \langle \boldsymbol{x} | \boldsymbol{y} \rangle$, *pd* to a polynomial kernel $K(\boldsymbol{x}, \boldsymbol{y}) = (1 + \langle \boldsymbol{x} | \boldsymbol{y} \rangle)^d$ with $d = 1, \ldots, 8$ and $r\gamma$ to a radial basis function $K(\boldsymbol{x}, \boldsymbol{y}) = \exp(-\gamma \|\boldsymbol{x} - \boldsymbol{y}\|^2)$ with $\gamma = 0.1, 1$ and 10 respectively.

from using a linear kernel whereas for LLE a polynomial kernel of degree 2 gives

the best results [2]. As far as the classification error is concerned, in the case of PCA, RBF kernels are at chance level and polynomial kernels of odd degree seem to be best. The error curve exhibits an instability for increasing degrees of the polynomial function. For LLE, on the other hand, the curve is smoother. Nonetheless, as data reduction method PCA clearly outperforms LLE in terms of classification error and data compression.

4.2 Original versus Processed Database

Here we evaluate classification performance for the processed and the original database for PCA and LLE using at the optimal settings from the previous section. Results are summarized in the following table, all values being averaged across illumination and perspective:

	PCA class. error	PCA $\sharp(SVs)$	LLE class. error	LLE $\sharp(SVs)$
original MPI	$5.16 \pm 2.18\%$	46 ± 4	$10.23 \pm 2.72\%$	120 ± 0
processed MPI	$6.59 \pm 2.60\%$	61 ± 4	$19.80 \pm 3.69\%$	120 ± 0

The superior classification performance for the original database confirms the need for the"clean up" processing applied to the MPI head database: the SVM used some of the obvious, but artifactual, cues for classification such as brightness and size. Note that in the case of LLE the number of SVs is constant for both databases but this is a ceiling effect: all the elements of the dataset are SVs, indicating that LLE may not be suited as a pre-processing algorithm for faces. LLE seems to be more sensitive to the "clean up" of the database suggesting that it may rely more strongly on obvious cues such as brightness or size. Again, the results of these simulations show that LLE performs poorly relative to PCA, both for classification and data compression. Since by definition LLE preserves local neighborhoods, the data is more difficult to be separated unless already *a priori* separable (what appears here not to be the case). PCA on the other hand finds the directions of main variance in the data therefore separating the data and doing an efficient preprocessing for classification. This may explain why LLE is less adapted for classification than PCA, at least for the face database under consideration.

4.3 Behavior with Respect to Illumination and Perspective

Here we assess the stability of the SVM with respect to changes in illumination (values averaged across perspectives) and perspective (values averaged along illuminations) of the processed database. The results are presented in figure 5 using the optimal parameter settings in each case. Classification performance

[2] We tried 5, 10 and 15 nearest neighbors for LLE using c_{opt} and found only very slight differences in performance for 10 and 15, 5 being clearly worse. In the following we always use the best, i.e. 15 nearest neighbors.

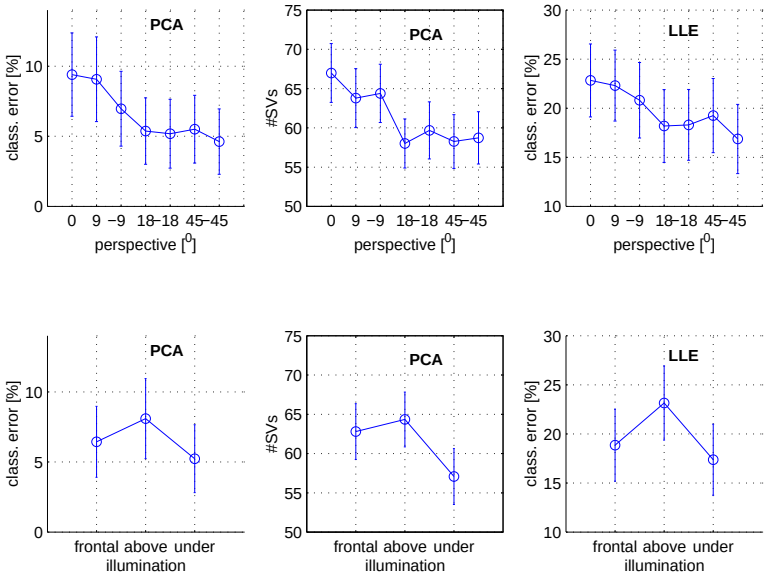

Fig. 5. Classification performance with respect to perspective and illumination for PCA and LLE (for LLE the number of SVs is not plotted since it is constant at 120).

as a function of orientation reveals a decrease of the classification error when moving away from a frontal perspective, i.e. classification is easier for non-frontal perspectives. This result holds for both PCA and LLE. Orientations of ± 18 and ± 45 seem largely equivalent. Note that this result indicates that some of the gender differences must be contained in the depth-profile of faces, for example, nose length or head curvature in depth, which are lost in a frontal projection. Furthermore, human subjects in psychophysical experiments exhibit a similar pattern of performance: they, too, show improved face recognition and gender classification for non-frontal presentation (the so called "3/4 view advantage" [13]).

Classification error and the number of SVs obtained as a function of illumination may show a pattern different from that of humans. Humans tend to perform best under natural illumination conditions, i.e light from above. Both for PCA and LLE performance is, however, worst for this illumination. Note that this effect is very small albeit consistent across PCA and LLE. A larger set of illumination conditions would be required to reach more definite conclusions on this issue.

5 Conclusions

The main results of the present study are, first, that PCA face space is clearly superior to that induced by LLE for classification tasks. Second, PCA face space

is linearly separable with respect to gender. Having a linear output stage has recently become a topic of interest in the context of complex, dynamical systems ("echo state" recurrent neural networks [14] and "liquid state machines" [15]) as it allows learning in such systems. As suggested in [15], this may even be a generic working principle of the brain to attempt to transform the problem at hand such that it becomes linearly separable. LLE, on the other hand, seems to require a polynomial kernel of degree two, forfeiting linear separability. For the poor performance of LLE compared to PCA there is, as in the case of the orientation dependency of classification, yet again an interesting parallel to human vision. It has been claimed that human expertise in face recognition during development from children to adults is brought about at least in part by a change in processing strategy: children focus on details (e.g. eyes or nose) whereas adults look at the whole face (sometimes referred to as "holistic processing") [16]. Eigenfaces in PCA face space are certainly fairly global (or holistic). Despite the fact that LLE face space is more homogeneous, as shown in sec. 3, and despite the algorithm displaying some biologically interesting properties like translation and rotation invariance, our results suggest that it is not well suited for gender classification. Finally we showed that the MPI head database contains factors such as size and lightness which are correlated with the classification result (as shown in sec. 4.2) but which cannot necessarily be relied upon to be informative either in real life or in other test sets for which the machine might be applied. Hence the database needs to be "cleaned up" (size and brightness normalization, centering) before it is useful for machine learning. This is an important issue also for other databases. Future work will focus on including additional biologically-motivated pre-processing techniques such as non-negative matrix factorization [17].

References

1. A.J. O'Toole, K.A. Deffenbacher, D. Valentin, K. McKee, D. Huff and H. Abdi. The Perception of Face Gender: the Role of Stimulus Structure in Recognition and Classification. *Memory & Cognition*, 26(1), 1998.
2. B. Moghaddam and M.-H. Yang. Gender Classification with Support Vector Machines. *Proceedings of the International Conference on Automatic Face and Gesture Recognition* (FG), 2000.
3. V. Blanz and T. Vetter. A Morphable Model for the Synthesis of 3D Faces. *Proc. Siggraph99*, pp. 187-194. Los Angeles: ACM Press, 1999.
4. R. O. Duda and P.E. Hart and D.G. Stork. *Pattern Classification*. John Wiley & Sons, 2001.
5. L. Sirovich, and M. Kirby. Low-Dimensional Procedure for the Characterization of Human Faces. *Journal of the Optical Society of America A*, 4(3), 519-24, 1987.
6. M. Turk and A. Pentland. Eigenfaces for Recognition. *Journal of Cognitive Neuroscience*, 3(1), 71-86, 1991.
7. A.J. O'Toole, H. Abdi, K.A. Deffenbacher and D. Valentin. Low-Dimensional Representation of Faces in Higher Dimensions of the Face Space. *Journal of the Optical Society of America A*, 10(3), 405-11, 1993.
8. S. T. Roweis and L.K. Saul. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science*, 290, 2000.

9. L.K. Saul and S. T. Roweis. An Introduction to Locally Linear Embedding. Report at AT&T Labs - Research, 2000.
10. B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. Smola and R.C. Williamson. Estimating the Support of a High-Dimensional Distribution. *Neural Computation*, 13(7), 2001.
11. V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.
12. A. B. A. Graf and S. Borer. Normalization in Support Vector Machines. *Proceedings of the DAGM*, LNCS 2191, 2001.
13. V. Bruce, T. Valentine and A.D. Baddeley. The Basis of the 3/4 View Advantage in Face Recognition. *Applied Cognitive Psychology*, 1:109-120, 1987.
14. H. Jaeger, The "Echo State" Approach to Analysing and Training Recurrent Neural Networks. GMD Report 148, German National Research Center for Information Technology, 2001.
15. W. Maass, T. Natschläger, and H. Markram. Real-Time Computing without Stable States: A New Framework for Neural Computation Based on Perturbations. *Neural Computation*, 2002 (in press).
16. M. Baenninger. The Development of Face Recognition: Featural or Configurational Processing? *Journal of Experimental Child Psychology*, 57(3), 377-96, 1994.
17. D. D. Lee and H. S. Seung. Learning the Parts of Objects by Non-Negative Matrix Factorization. *Nature*, 401:788-791, 1999.