Structured Covariance Matrices for Statistical Image Object Recognition

J. Dahmen, D. Keysers, M. Pitz, H. Ney

Lehrstuhl für Informatik VI, Computer Science Department
RWTH Aachen - University of Technology
D-52056 Aachen, Germany
{dahmen, keysers, pitz, ney}@informatik.rwth-aachen.de

Abstract. In this paper we present different approaches to structuring covariance matrices within statistical classifiers. This is motivated by the fact that the use of full covariance matrices is infeasible in many applications. On the one hand, this is due to the high number of model parameters that have to be estimated, on the other hand the computational complexity of a classifier based on full covariance matrices is very high. We propose the use of diagonal and band-matrices to replace full covariance matrices and we also show that computation of tangent distance is equivalent to using a structured covariance matrix within a statistical classifier.

1 Introduction

In the last few years, the use of Bayesian classifiers based on Gaussian mixture densities or kernel densities proved to be very efficient for many pattern recognition tasks, among them speech recognition, machine translation and object recognition in images [1, 2, 3, 7]. One drawback of this approach is the fact that the number of model parameters for such a classifier is extremely high, requiring a very large amount of training data (which is not always available) for reliable parameter estimation. A common approach to overcome this difficulty is the use of diagonal instead of full covariance matrices, i.e. the use of variance vectors. In this paper we investigate other possibilities to structure covariance matrices (the variance vector being a very simple structuring approach). On the one hand, we will do so by assuming that the grayvalue of a certain pixel only depends on the grayvalues of the neighbouring pixels. We will also show that computation of Simard's tangent distance [14] can be interpreted as a special structure of covariance matrices within a statistical classifier.

In the next Section, we will briefly describe the US Postal Service database (USPS) which we used to carry out our experiments. Before discussing possible approaches to structuring covariance matrices in Section 4, we will describe the statistical classifier used in our experiments in Section 3. After presenting experimental results in Section 5 (as well as a comparison of our results with those reported by other international research groups), we will conclude the paper in Section 6.

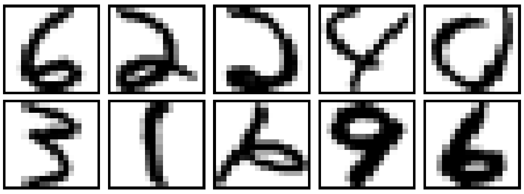

Fig. 1. Example images taken from the USPS database

2 The US Postal Service Database

The USPS database (`ftp://ftp.kyb.tuebingen.mpg.de/pub/bs/data/`) is a well known handwritten digit recognition database. It contains 7291 training objects and 2007 test objects. The digits are isolated and represented by a 16×16 pixels sized grayscale image (see Figure 1). Making use of *appearance based pattern recognition* in our experiments, we interpret each pixel as a feature, obtaining a 256-dimensional feature vector. The USPS recognition task is known to be very hard, with a human error rate of about 2.5% on the testing data [14]. An advantage of the USPS task is the availability of many recognition results reported by international research groups, allowing for a fair comparison of results.

3 The Statistical Classifier

To classify an observation $x \in \mathbb{R}^D$ we use the Bayesian decision rule [6, pp. 10-39]

$$x \longmapsto r(x) = \underset{k}{\operatorname{argmax}} \{p(k)p(x|k)\} \tag{1}$$

where $p(k)$ is the prior probability of class k, $p(x|k)$ is the class conditional probability for the observation x given class k and $r(x)$ is the decision of the classifier. As neither $p(k)$ nor $p(x|k)$ are known, we have to choose models for the respective distributions and estimate their parameters using the training data.

3.1 Gaussian Mixture Densities

In our experiments, we set $p(k) = \frac{1}{K}$ for each class k (as it is not obvious why a certain digit should have a higher prior probability than another) and model $p(x|k)$ by using Gaussian mixture densities or kernel densities respectively. A Gaussian mixture is defined as a linear combination of Gaussian component densities $\mathcal{N}(x|\mu_{ki}, \Sigma_{ki})$, leading to the following expression for the class conditional probabilities:

$$p(x|k) = \sum_{i=1}^{I_k} c_{ki} \cdot \mathcal{N}(x|\mu_{ki}, \Sigma_{ki}) \tag{2}$$

where I_k is the number of component densities used to model class k, c_{ki} are weight coefficients (with $c_{ki} > 0$ and $\sum_{i=1}^{I_k} c_{ki} = 1$, which is necessary to ensure that $p(x|k)$ is a probability density function), μ_{ki} is the mean vector and Σ_{ki} is the covariance matrix of component density i of class k.

Maximum-likelihood parameter estimation can now be done using the Expectation-Maximization algorithm [4] in combination with a Linde-Buzo-Gray based clustering procedure [10]. More information on that topic (for diagonal covariance matrices) can be found in [2] or [1] respectively.

3.2 Kernel Densities

In the case of kernel densities (also called parzen windows or parzen densities) [5, pp. 147-153], each training sample x_n defines a Gaussian single density $\mathcal{N}(x|x_n, \Sigma_{x_n})$ with an estimated covariance matrix Σ_{x_n}, that is the sample itself is interpreted as mean vector. Thus, kernel densities might be interpreted as an extreme case of a mixture density model.

To classify an observation x, we now use the decision function

$$x \longmapsto r(x) = \underset{k}{\operatorname{argmax}} \{p_{KD}(x|k)\}, \quad \text{where} \tag{3}$$

$$p_{KD}(x|k) = \frac{1}{N_k} \sum_{n=1}^{N_k} \mathcal{N}(x|x_n, \Sigma_{x_n}) \tag{4}$$

and N_k is the number of training samples belonging to class k.

A typical problem for statistical classifiers based on the models described above is the estimation of covariance matrices. In case of the USPS task, with feature vectors $x \in \mathbb{R}^{256}$, a single covariance matrix requires (due to symmetries) the estimation of $256 \cdot (256+1)/2 = 32.896$ parameters. Given only 7.291 training samples, this is infeasible. A common approach to overcome this difficulty is the use of variance pooling

- *class specific variance pooling* :
  estimate only a single Σ_k for each class k, i.e. $\Sigma_{ki} = \Sigma_k \;\; \forall\, i = 1, ..., I_k$
- *global variance pooling* :
  estimate only a single Σ, i.e. $\Sigma_{ki} = \Sigma \;\; \forall\, k = 1, ..., K$ and $\forall\, i = 1, ..., I_k$

in combination with diagonal covariance matrices, i.e. variance vectors. In our kernel density experiments, we made use of class specific variance pooling, that is we computed the empirical covariance matrix Σ_k for each class k and set $\Sigma_{x_n} := \Sigma_k$ for each observation x_n of class k. In contrast to this, our mixture density based experiments were conducted using globally pooled variances, as this proved to be the best choice.

Note that the use of a diagonal covariance matrix can be interpreted as a very simple approach to structuring covariance matrices, where a rather harsh approximation of a full covariance matrix is used in order to reduce the number of free model parameters. In the following Section, we will present alternative, more sophisticated approaches to this problem.

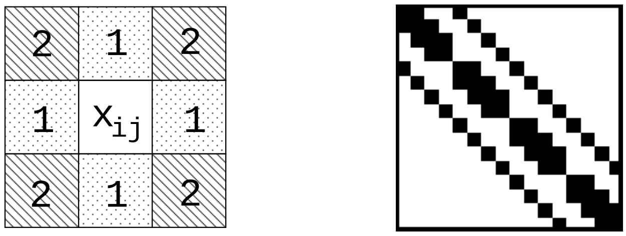

Fig. 2. Neigbourhoods N_1 (1), N_2 (1, 2) used (left). Resulting band structure of the inverse covariance matrix Σ^{-1} for N_1 and 4×4 pixels sized images (right). Black pixels represent non-zero entries in Σ^{-1}.

4 Structuring Covariance Matrices

In this section we will present two approaches to estimating structured covariance matrices for image object recognition. The first is based on pixel neighbourhoods and their influence on the covariance matrix Σ, the second is derived from a probabilistic interpretation of tangent distance.

4.1 Structures based on Pixel Neighbourhoods

Using full covariance matrices for object recognition implies the possibility that any two pixels within an image are correlated. On the other hand, using diagonal covariance matrices, we assume that there is no correlation between different pixels at all. Both such approaches are somewhat extreme: the first suffers from a large amount of parameters, whereas the latter may be an unrealistic model in some applications. As a compromise, one could use a full covariance matrix with the restriction that the grayvalue of a given pixel only depends on the grayvalues of its neighbours. Thus, the number of non-zero entries in the respective inverse covariance matrix can be significantly reduced.

Regarding the neighbourhoods N_1 and N_2 as shown in Figure 2 and assuming that the grayvalue of a pixel x_{ij} only depends on its neighbouring pixels, the respective inverse covariance matrix Σ^{-1} has a band structure (this can be shown using Markov random field theory [9]), with the number of bands increasing as the regarded neighbourhood grows (four bands for N_1, eight for N_2). Thus, any entry of Σ^{-1} that does not lie on the diagonal or the bands is zero. Note that some entries on the first band are zero, too (cp. Figure 2). This is due to the fact that wrap-around is not considered, e.g. a pixel at the left border of an image is not a neighbour of the corresponding pixel at the right border.

Considering this, a maximum-likelihood estimation of Σ_k (i.e. maximization of $\prod_{n=1}^{N_k} p(x_{nk}|k)$ with respect to Σ_k, given the training observations x_{nk}, $n = 1, ..., N_k$) yields the interesting result, that we can only give estimations for those entries in Σ_k that lie on the diagonal or the bands. Thus, we know each entry in Σ_k that we do not know in ${\Sigma_k}^{-1}$ (where we have knowledge about the occurences of zeros) and vice versa. Hence, an estimation for ${\Sigma_k}^{-1}$ (under the constraint that only neighbouring pixel depend on each other) can be found by solving the bilinear equation system

$$\Sigma_k \cdot \Sigma_k{}^{-1} = I \tag{5}$$

where I is the matrix of identity. With $\Sigma_k, \Sigma_k{}^{-1} \in \mathbb{R}^{D\times D}$, this yields D^2 equations with D^2 unknowns. In our experiments, the solution of this equation system is obtained by applying the Gauss-Seidel algorithm [11, pp. 864-869].

4.2 A Structure based on Tangent Distance

In 1993, Simard et al. proposed an invariant distance measure called *tangent distance*, which proved to be especially effective for optical character recognition [14]. The authors observed that reasonably small transformations of certain objects (like digits) do not affect class-membership. Simple distance measures like the Euclidean distance do not account for this, instead they are very sensitive to transformations like scaling, translation, rotation or axis deformations. When an image x of size $I \times J$ is transformed (e.g. scaled and rotated) with a transformation $t(x, \alpha)$ which depends on L parameters $\alpha \in \mathbb{R}^L$ (e.g. the scaling factor and the rotation angle), the set of all transformed images

$$M_x = \{t(x, \alpha) : \alpha \in \mathbb{R}^L\} \subset \mathbb{R}^{I\times J} \tag{6}$$

is a manifold of at most L dimensions. The distance between two images can now be defined as the minimum distance between their according manifolds, being truly invariant with respect to the L transformations regarded. Unfortunately, computation of this distance is a hard optimization problem and the manifolds needed have no analytic expression in general. Therefore, small transformations of an image x are approximated by a tangent subspace $\hat{M}_x$ to the manifold M_x at the point x. Those transformations can be obtained by adding to x a linear combination of the vectors $T_l(x), l = 1, ..., L$ that span the tangent subspace. Thus, we obtain as a first-order approximation of M_x:

$$\hat{M}_x = \{x + \sum_{l=1}^{L} \alpha_l \cdot T_l(x) \ : \alpha \in \mathbb{R}^L\} \subset \mathbb{R}^{I\times J} \tag{7}$$

Now, the single sided tangent distance $D_T(x, \mu)$ between an image x and a reference image μ is defined as

$$D_T(x, \mu) = \min_{\alpha}\{\|x + \sum_{l=1}^{L} \alpha_l \cdot T_l(x) - \mu\|^2\} \tag{8}$$

The *tangent vectors* $T_l(x)$ can be computed using finite differences between the original image x and a small transformation of x [14]. A double sided TD can also be defined by approximating M_x and M_μ and minimizing the distance over all possible combinations of the respective parameters. In our experiments, we computed the seven tangent vectors for translations (2), rotation, scaling, axis deformations (2) and line thickness, as proposed by Simard. Assuming that the

Table 1. Error rates on USPS for varying image sizes and structures of Σ.

Image Size	Structure	Error Rate [%]
8×8	Diagonal	5.7
	Band using N_1	5.5
	Band using N_2	5.1
	full	4.6
	tangent structure	4.6
16×16	diagonal + tangent, GMD	2.7
	diagonal + tangent, KD	2.2

tangent vectors are orthogonal (which can be achieved using a singular value decomposition), Eq. (8) can be solved efficiently by computing

$$D_T(x, \mu) = \|x - \mu\|^2 - \sum_{l=1}^{L} \frac{[(x-\mu)^t \cdot T_l(x)]^2}{\|T_l(x)\|^2} \tag{9}$$

Conceptionally, tangent variations of the references can be incorporated into a statistical classifier by modeling Gaussian normal distributions via $\mathcal{N}(x|\mu + \sum_{l=1}^{L} \alpha_l \cdot T_l(\mu), \Sigma)$ with unknown α_l. If we further assume a Gaussian distribution for the parameter set α with zero mean and variance approaching infinity, one can show that this probability can be computed using the following structured covariance matrix:

$$\hat{\Sigma} := \lim_{\kappa \to \infty} \left(\Sigma + \kappa \sum_{l=1}^{L} \frac{T_l(\mu) T_l(\mu)^t}{T_l(\mu)^t \Sigma^{-1} T_l(\mu)} \right) \tag{10}$$

where Σ is the empirical covariance matrix of the data. With κ approaching infinity, variances along the directions in feature space defined by the tangent vectors approach infinity, too. Thus, variations of the reference images along these directions are not considered. Note that the matrix $\hat{\Sigma}$ cannot be used explicitly (as it does not exist for $\kappa \to \infty$), yet calculating single-sided tangent distance is equivalent to using $\hat{\Sigma}$. As the required calculations to prove this statement are rather lengthy, they are omitted here. A detailed discussion of this topic can be found in [8].

5 Results

We started our experiments by applying the kernel density based classifier to the USPS task. As the solution of the bilinear equation system (5) is very time consuming, the USPS images were scaled down to a size of 8×8 pixels. Experiments were done using the following structures for the (class specifically pooled) covariance matrices: (a) diagonal, (b) band structure using N_1 or N_2 respectively, (c) structure via tangent distance as shown in Eq. (10) and (d) full covariance matrix. The results obtained are shown in Table 1. As one would have expected, estimation of a band structured covariance matrix reduces the error rate as compared to a diagonal structure. Best results are obtained using a full covariance

Table 2. Results reported on USPS

Author	Method	Error [%]
Simard et al., 1993	Human Performance	2.5
Vapnik, 1995	Decision Tree C4.5	16.2
Vapnik, 1995	Two-Layer Neural Net	5.9
Simard et al., 1998	Five-Layer Neural Net	4.2
Schölkopf, 1997	Support Vectors	4.0
Schölkopf et al., 1998	Invariant Support Vectors	3.0
Simard et al., 1993	Tangent Distance	*2.6
This work:	Gaussian Mixtures + tangents	2.7
	Kernel Densities + tangents	2.2

*: 2400 machine printed digits were added to the training set

matrix, which is not surprising, since we only estimated a single covariance matrix per class, using downscaled USPS images. Interestingly, using the tangent distance based structure yields the same results as compared to a full covariance matrix, but - at the same time - reduces the computational complexity significantly. Using the original 16×16 pixels sized USPS data, the tangent structure (3.3%) significantly outperforms a full covariance matrix (6.3%, as the number of free parameters increases by a factor of 16).

We therefore embedded tangent distance into a Gaussian mixture density based classifier, based on diagonal, globally pooled covariance matrices. On the original 16×16 pixels sized USPS images, this yields an excellent error rate of 2.7% using double-sided tangent distance. Using a bagged kernel density based classifier, this error rate could be further reduced to 2.2%. These experiments were conducted on virtually augmented USPS data, where each image was shifted into the directions of the N_2-neighbourhood, yielding $9 \cdot 7291 = 65619$ training samples (using other transformations to create virtual data did not improve the error rate any further). A similar approach was used on the testing data, where the final decision for the original test sample was achieved by using the sum rule. Detailed information on the use of virtual data within statistical classifiers and its impact on the classification error rate can be found in [1, 7]. Note that using virtual data in combination with tangent distance is useful, as the shifted images lead to a better approximation of the true manifolds (tangent distance only approximates image shifts). A comparison of our results with that reported by other groups can be found in Table 2, proving them to be state-of the art.

6 Conclusions

In this paper we presented a novel approach to using structured covariance matrices for image object recognition within a statistical classifier. The structures we proposed are based on a neighbourhood concept (only neighbouring pixels depend on each other) and on a probabilistic interpretation of Simard's tangent distance. Using such structures, the number of model parameters that have to be estimated can be considerably reduced. The advantage of this reduction is twofold: On the one hand, parameter estimation is more reliable, on the other hand the computational complexity of the classifier is reduced. We obtained excellent results on the US Postal Service handwritten digit recognition task, especially when using tangent distance to structure the respective covariance matrices (2.2% error rate using a kernel density based classifier and virtual data).

References

1. J. Dahmen, D. Keysers, M. Güld, H. Ney, "Invariant Image Object Recognition using Gaussian Mixture Densities", Proceedings of the 15th International Conference on Pattern Recognition, Barcelona, Spain, September 2000, in press.
2. J. Dahmen, K. Beulen, M. Güld, H. Ney, "A Mixture Density Based Approach to Object Recognition for Image Retrieval", Proceedings of the 6th International RIAO Conference on Content-Based Multimedia Information Access, Paris, France, April 2000, in press.
3. J. Dahmen, R. Schlüter, H. Ney, "Discriminative Training of Gaussian Mixtures for Image Object Recognition", in W. Förstner, J. Buhmann, A. Faber, P. Faber (eds.): Proceedings of the 21. Symposium of the German Association for Pattern Recognition (DAGM), Bonn, Germany, pp. 205-212, September 1999.
4. A.P. Dempster, N.M. Laird, D.B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society*, 39(B), pp. 1-38, 1977.
5. L. Devroye, L. Györfi, G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, Springer, New York, 1996.
6. R. O. Duda, P. E. Hart, *Pattern Classification and Scene Analysis*, John Wiley & Sons, 1973.
7. D. Keysers, J. Dahmen, T. Theiner, H. Ney, "Experiments with an Extended Tangent Distance", Proceedings of the 15th International Conference on Pattern Recognition, Barcelona, Spain, September 2000, in press.
8. D. Keysers, J. Dahmen, H. Ney, "A Probabilistic View on Tangent Distance", Proceedings of the 22. Symposium of the German Association for Pattern Recognition (DAGM), Kiel, Germany, September 2000, this volume.
9. S. Z. Li, Markow Random Field Modelling in Computer Vision, Springer, Tokyo, Japan, 1995.
10. Y. Linde, A. Buzo und R. M. Gray, "An Algorithm for Vector Quantizer Design," *IEEE Transactions on Communications*, Vol. 28, No. 1, pp. 84-95, 1980.
11. W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery, *Numerical Recipes in C*, University Press, Cambridge, 1992.
12. B. Schölkopf, *Support Vector Learning*, Oldenbourg Verlag, Munich, 1997.
13. B. Schölkopf, P. Simard, A. Smola, V. Vapnik, "Prior Knowledge in Support Vector Kernels," M. Jordan, M. Kearns, S. Solla (eds.): *Advances in Neural Information Processing Systems 10*, MIT Press, pp. 640-646, 1998.
14. P. Simard, Y. Le Cun, J. Denker, "Efficient Pattern Recognition Using a New Transformation Distance," S.J. Hanson, J.D. Cowan, C.L. Giles (eds.): *Advances in Neural Information Processing Systems 5*, Morgan Kaufmann, San Mateo CA, pp. 50-58, 1993.
15. V. Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York, pp.142-143, 1995.

This article was processed using the LaTeX macro package with LLNCS style