Invariant Image Object Recognition using Mixture Densities

Jörg Dahmen, Daniel Keysers, Mark Oliver Güld, Hermann Ney
Lehrstuhl für Informatik VI
RWTH Aachen - University of Technology
D-52056, Aachen, Germany
{dahmen, keysers, gueld, ney}@informatik.rwth-aachen.de

Abstract

In this paper we present a mixture density based approach to invariant image object recognition. We start our experiments using Gaussian mixture densities within a Bayesian classifier. Invariance to affine transformations is achieved by replacing the Euclidean distance with SIMARD*'s tangent distance. We propose an approach to estimating covariance matrices with respect to image invariances as well as a new classifier combination scheme, called the virtual test sample method. On the US Postal Service handwritten digits recognition task (USPS), we obtain an excellent classification error rate of* 2.7%*, using the original USPS training and test sets.*

1 Introduction

In this paper we present a mixture density (MD) based approach to invariant image object recognition. We propose a Gaussian mixture density (GMD) based Bayesian classifier and extend this non-invariant standard approach using SIMARD's tangent distance (TD) [1] instead of Euclidean distance. We also use TD for the reliable estimation of covariance matrices, which is especially important if only few training samples are available. Furthermore, we propose a new classifier combination scheme called the virtual test sample method (VTS). The effectiveness of our approach is shown by applying it to the widely used USPS recognition task. In the experiments, we make use of appearence based pattern recognition, that is we interpret each pixel of an image as a feature, optionally performing feature reduction by using a linear discriminant analysis (LDA) [2, pp.114-123]. Using VTS and LDA, the GMD standard approach yields a test error of 3.4%. The error rate can be improved to 2.7% by using TD in recognition and by estimating the proposed *tangent covariance matrix* (without LDA).

1.1 Related work

While appearance based image object recognition is common, the use of (invariant) statistical classifiers such as the one we propose is not. MOGHADDAM & PENTLAND used GMDs for view-based image recognition, accounting for invariances by assuming appropriate training samples and suitable image normalization [3]. SCHIELE employed histogram based image features within a Bayesian classifier, but did not use mixture densities to model the required probability densities [4]. HINTON et al. applied TD to define a modified version of a principal components analysis within a linear autoencoder based classifier [5]. This approach is similar to computing a maximum approximation within a MD based classifier. HASTIE et al. computed suitable prototype vectors from a given training set with respect to TD, which can be used to speed up nearest neighbour classification (by using a few prototype vectors instead of the possibly large training set) [6]. Many authors such as SCHWENK use TD within artificial neural nets [7]. Finally, the virtual test sample method proposed in Section 3 was motivated by KITTLER's research on classifier combination [8].

2 The GMD based standard approach

In the statistical GMD based 'standard' approach, we classify an observation $x \in \mathbb{R}^D$ using the Bayesian decision rule [2, pp.10-39]

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

where $p(k)$ is the *a priori* probability of class k, $p(x|k)$ is the *class conditional* probability for the observation x given class k and $r(x)$ is the classifier's decision. As neither $p(k)$ nor $p(x|k)$ are known, we have to choose models for them and estimate their parameters with the training data. As we are performing digit recognition in our experiments, we set $p(k) = \frac{1}{K}$ for each class k (given K classes) and model $p(x|k)$ by using GMDs, being a linear combination of Gaussian component densities $\mathcal{N}(x|\mu_{ki}, \Sigma_{ki})$

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

where I_k is the number of component densities used to model class k, c_{ki} are weight coefficients (with $c_{ki} > 0$ and $\sum_i c_{ki} = 1$), μ_{ki} is the mean vector and Σ_{ki} is the covariance matrix of component density i of class k. To avoid the problems of estimating a covariance matrix in a high dimensional feature space (cp. Sections 3 and 5), i.e. to keep the number of parameters to be estimated small, we make use of global covariance pooling in the experiments, that is we only estimate a single Σ, i.e. $\Sigma_{ki} = \Sigma$ $\forall\, k \in \{1, ..., K\}$ and $\forall\, i \in \{1, ..., I_k\}$. This model consistently outperformed class specific variance pooling as well as doing no pooling at all in our experiments. Furthermore, we only use a diagonal covariance matrix, i.e. a variance vector. This does not necessarily imply a loss of information, as a MD of that form can still approximate any density function with arbitrary precision. Optionally, we use an LDA for feature reduction. As performing an LDA on the original ten-class USPS data would only yield a maximum of nine features, we perform a cluster analysis on the training data first. Creating 40 clusters (using 39 LDA features), we obtained the best results in the experiments. Maximum-likelihood parameter estimation is then performed using the Expectation-Maximization (EM) algorithm. More information on this topic can be found in [9, 10].

3 Creating virtual data

A typical drawback of statistical classifiers is their need for a large amount of training data, which is not always available. To overcome this difficulty, we create virtual training data. The basic idea is to choose a transformation which respects class membership and to apply it to each training sample. In the experiments, we used ± 1 pixel shifts to create $9 \cdot 7291 = 65.619$ training samples of size 18×18 pixels from the original 7291 USPS training samples (of size 16×16 pixels). By doing so, parameter estimation is not only more reliable, but we also incorporate local invariances with respect to the chosen transformation(s) in our MD model.

3.1 The virtual test sample method

Similar to creating virtual training data, we propose the following virtual test sample method (VTS). Using our a-priori knowledge again, we create A virtual test samples $x_1, ..., x_A$ by applying a number of shifts to each test image (we use ± 1 pixel shifts, i.e. $A = 9$, other transformations might be considered in other domains). As an image cannot be shifted into different directions at the same time, we can create a final decision by computing

$$p(x|k) = \sum_{\alpha=1}^{A} p(\alpha) \cdot p(x|\alpha, k) = \frac{1}{A} \sum_{\alpha=1}^{A} p(x_\alpha|k) \qquad (3)$$

(assuming that the a-priori probabilities $p(\alpha)$ are equal for all transformations considered). As the term $1/A$ does not depend on k, we may neglect it for classification purposes. Note that this motivation for the sum rule differs from that proposed by KITTLER in [8]. Using multiple classifiers to classify a single test pattern, he assumed that the a-posteriori probabilities computed by the respective classifiers do not differ much from the a-priori probabilities to justify the sum rule. In contrast to this, using multiple test patterns and a single classifier, Eq. (3) simply follows from the fact that the transformations considered are mutually exclusive. The key idea behind VTS is that we are able to use classifier combination schemes and their benefits without having to create multiple classifiers. Instead, we simply create virtual test samples. Thus, classifying a pattern using VTS has the same computational complexity as using any other combination scheme, but the training phase remains unaffected. Despite its simplicity, VTS proved to be very effective in our experiments.

4 Overview of 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 [1]. 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

$$M_x = \{t(x, \alpha) : \alpha \in \mathbb{R}^L\} \subset \mathbb{R}^{I \times J} \qquad (4)$$

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:

$$\hat{M}_x = \{x + \sum_{l=1}^{L} \alpha_l \cdot T_l(x) \ : \alpha \in \mathbb{R}^L\} \subset \mathbb{R}^{I \times J} \qquad (5)$$

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

$$D_T(x, \mu) = \min_{\alpha}\{||x + \sum_{l=1}^{L} \alpha_l \cdot T_l(x) - \mu||^2\} \qquad (6)$$

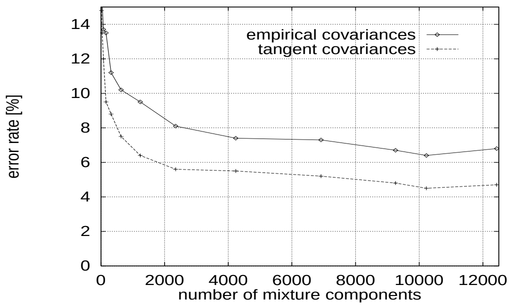

Figure 1. Empirical variance vs. tangent variance: error rates with respect to total number of mixture components (9-1, no LDA)

The tangent vectors $T_l(x)$ can be computed using simple finite differences between the original image x and a small transformation of x [1]. 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 the experiments, we computed the seven tangent vectors for translations (2), rotation, scaling, axis deformations (2) and line thickness, as proposed by Simard [1]. Assuming that the tangent vectors are orthogonal (which can be achieved using a singular value decompostion), Eq. (6) can be solved efficiently by computing

$$D_T(x, \mu) = ||x - \mu||^2 - \sum_{l=1}^{L} \frac{[(x-\mu)^t \cdot T_l(x)]^2}{||T_l(x)||^2} \quad (7)$$

5 Parameter estimation with TD

Instead of computing the empirical covariance matrix Σ of the given training samples, we can use Eq. (5) to implicitly create an infinite amount of training samples and compute the respective tangent covariance matrix Σ_T, which should be a better estimate for the covariance:

$$\Sigma_T = \frac{1}{N} \int p(\alpha) \cdot \sum_{n=1}^{N} (x_{n,\alpha} - \mu)(x_{n,\alpha} - \mu)^t \, d\alpha \quad (8)$$

where $x_{n,\alpha} = x_n + \sum_{l=1}^{L} \alpha_l \cdot T_l(x_n)$ is a local transformation of the n-th training pattern, N is the number of training samples with mean μ and $p(\alpha)$ is the distribution of the parameters α. With $\int p(\alpha)\, d\alpha = 1$, $E(\alpha) = 0$ and some elementary calculations, Eq. (8) can be written as

$$\Sigma_T = \Sigma + \frac{1}{N} \sum_{n=1}^{N} T_{x_n} \Sigma_\alpha T_{x_n}^t \quad (9)$$

with $T_{x_n} \in \mathbb{R}^{D \times L}$ being the matrix representation of the tangent vectors of training sample x_n and $\Sigma_\alpha \in \mathbb{R}^{L \times L}$ the covariance matrix of parameters α (we use $\Sigma_\alpha = I$ in our experiments). Note that

$$\mu_T = \frac{1}{N} \int p(\alpha) \cdot \sum_{n=1}^{N} x_{n,\alpha} \, d\alpha = \mu \quad (10)$$

that is, the empirical sample mean μ does not change in the presence of tangent vectors. In the experiments we will show that combining the explicit creation of virtual training data with this implicit approach is advisable.

Table 1. USPS results with varying variance estimation and distance measures, with and without LDA

Method:	Error rate [%]			
	1-1	1-9	9-1	9-9
GMD	8.0	6.6	6.4	6.0
GMD, LDA	6.7	5.9	4.5	3.4
MD, Σ_T, Euclidean	6.4	4.8	4.5	4.3
MD, Σ_T, tangent	3.9	3.6	3.4	2.9

6 Results

We started our experiments by applying the GMD based standard approach to USPS. Table 1 shows the achieved results with and without LDA feature reduction. The notation '*a*-*b*' indicates, that we increased the number of training samples by a factor of a and that of the test samples by a factor of b. Thus, b=9 indicates that we performed VTS as proposed in Section 3. Note that by using the LDA, the error rate drops from 6.0% to 3.4%. This is mainly due to the problem of estimating variances in a high dimensional feature space, as the next experiment shows. In this experiment, we used Eq. (9) to estimate variances in the EM training phase without doing a feature reduction. Surprisingly, by simply computing the tangent variances, the error rate drops from 6.0% to 4.3%. A comparison of both approaches with respect to the total number of densities used in the probabilistic model can be found in Figure 1. Apparently, computing tangent variances in combination with explicitly creating virtual training data is a good means to overcome the difficulties in estimating a covariance matrix in a high dimensional feature space.

In another experiment, we replaced the Euclidean distance used in the Gaussian component densities by the single sided TD in the recognition step, whereas the training step was still performed using Euclidean distance, further reducing the error rate from 4.3% to 2.9%. The results of these experiments are shown in Table 1. The best result of 2.9% could be further reduced to 2.7% by calculating the double sided TD in recognition (using a total of about 10.000 mixture components, i.e. on average about 1000 per class). We were not able to obtain a result better than 3.0% error without using tangent variances, but using a bagged kernel density based classifier reduced the error rate to 2.2% [14]. A comparison of our USPS results with that reported by other groups can be found in Table 2, proving them to be excellent. Other groups report results of 2.6% error, but these were achieved by by adding about 2.500 machine

Table 2. Experimental results on USPS

Method		Error Rate [%]
Human Performance [1]		2.5
Decision Tree C4.5 [11]		16.2
Two-Layer Neural Net [11]		5.9
5-Layer Neural Net (LeNet1) [12]		4.2
Invariant Support Vectors [13]		3.0
This work:	GMD	4.5
	GMD, VTS	3.4
	MD, VTS, TD	2.7

printed digits to the training set [1, 15]. We also performed experiments with Fourier-transformation based invariants, invariant moments and discriminative training of Gaussian mixtures [10], yet so far none of these approaches could improve our best result. Furthermore, using TD in the training phase yielded no improvement. We also used AdaBoost [16] to boost our GMD classifier, using LDA reduced features. We were able to reduce the 9-1 error rate from 4.5% to 4.2%, yet VTS (reducing the error rate from 4.5% to 3.4%) outperformed boosting on this particular task. As for computational complexity, the standard GMD approach is cheap, requiring less than 0.1 CPU seconds to classify a single pattern (39 LDA features) on a Digital Alpha 500 MHz CPU. Using single sided TD (no LDA) takes about 1 CPU second and the computationally expensive double sided TD requires about 50 CPU seconds. Considering error rate vs. computational complexity, single sided TD might be considered the best choice for use in practice.

7 Conclusion

In this paper we presented an invariant mixture density based approach to object recognition, obtaining an excellent error rate of 2.7% on the USPS test set, using the original training set. Incorporating tangent vectors into the EM training (to compute the proposed tangent variances) and tangent distance itself into the recognition phase (replacing Euclidean distance) proved to be very effective, especially when combined with the creation of virtual training data and the proposed virtual test sample method. Besides applying the proposed methods to LDA reduced features, we are currently working on using the training data to improve the estimate of the covariance matrix Σ_α of the transformation parameters α (see Section 5).

References

[1] P. Simard, Y. Le Cun, and J. Denker. Efficient pattern recognition using a new transformation distance. In S. Hanson, J. Cowan, and C. Giles, editors, *Advances in Neural Information Processing Systems 5*, Morgan Kaufmann, San Mateo CA, pp. 50-58, 1993.

[2] R. Duda and P. Hart. *Pattern classification and scene analysis*. John Wiley & Sons, 1973.

[3] B. Moghaddam and A. Pentland. Probabilistic visual learning for object representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 19, No. 7, pp. 696-710, July 1997.

[4] B. Schiele and J. Crowley. Probabilistic object recognition using multidimensional receptive field histograms. *Procs. 13th Int. Conf. on Pattern Recognition*, Vienna, Austria, pp. 50-54, 1996.

[5] G. Hinton, M. Revow, and P. Dayan. Recognizing handwritten digits using a mixture of linear models. In G. Tesauro, D. Touretzky and T. Leen, editors, *Advances in Neural Information Processing Systems 7*, Morgan Kaufmann, San Mateo CA, pp. 1015-1022, 1995.

[6] T. Hastie, P. Simard, and E. Säckinger. Learning prototype models for tangent distance. In G. Tesauro, D. Touretzky and T. Leen, editors, *Advances in Neural Information Processing Systems 7*, Morgan Kaufmann, San Mateo CA, pp. 999-1006, 1995.

[7] H. Schwenk and M. Milgram. Transformation invariant autoassociation with application to handwritten character recognition. In G. Tesauro, D. Touretzky and T. Leen, editors, *Advances in Neural Information Processing Systems 7*, Morgan Kaufmann, San Mateo CA, pp. 991-998, 1995.

[8] J. Kittler, M. Hatef, and R. Duin. Combining classifiers. *Procs. 13th Int. Conf. on Pattern Recognition*, Vienna, Austria, pp. 897-901, 1996.

[9] J. Dahmen, K. Beulen, M. Güld, and H. Ney. A mixture density based approach to object recognition for image retrieval. *Procs. 6th International RIAO Conference on Content-Based Multimedia Information Access*, Paris, France, pp. 1632-1647, April 2000.

[10] J. Dahmen, R. Schlüter, and H. Ney. Discriminative training of Gaussian mixtures for image object recognition. *Procs. 21. Symposium German Association for Pattern Recognition (DAGM)*, pp. 205-212, Bonn, Germany, September 1999.

[11] V. Vapnik. *The nature of statistical learning theory*. Springer, New York, pp. 142-143, 1995.

[12] P. Simard, Y. Le Cun, J. Denker, and B. Victorri. Transformation invariance in pattern recognition — tangent distance and tangent propagation. *Lecture Notes in Computer Science*, Vol. 1524, Springer, pp. 239-274, 1998.

[13] B. Schölkopf, P. Simard, A. Smola, and V. Vapnik. Prior knowledge in support vector kernels. M. Jordan, M. Kearns and S. Solla, editors, *Advances in Neural Information Processing Systems 10*, MIT Press, pp. 640-646, 1998.

[14] D. Keysers, J. Dahmen, T. Theiner, and H. Ney. Experiments with an extended tangent distance. *15th International Conference on Pattern Recognition*, Barcelona, Spain, September 2000, this volume.

[15] H. Drucker, R. Schapire, and P. Simard. Boosting performance in neural networks. *International Journal of Pattern Recognition and Artificial Intelligence*, Vol. 7, No. 4, pp. 705-719, 1993.

[16] Y. Freund and R. Schapire. Experiments with a new boosting algorithm. *Procs. 13th International Conference on Machine Learning*, Bari, Italy, July 1996.