Statistical Image Object Recognition using Mixture Densities

Jörg Dahmen (`dahmen@informatik.rwth-aachen.de`), Daniel
Keysers, Hermann Ney, and Mark Oliver Güld
Lehrstuhl für Informatik VI, Computer Science Department,
RWTH Aachen - University of Technology,
D-52056 Aachen, Germany

November 13, 2000

Abstract. In this paper, we present a mixture density based approach to invariant image object recognition. To allow for a reliable estimation of the mixture parameters, the dimensionality of the feature space is optionally reduced by applying a robust variant of linear discriminant analysis. Invariance to affine transformations is achieved by incorporating invariant distance measures such as tangent distance. We propose an approach to estimating covariance matrices with respect to image variabilities as well as a new approach to combined classification, called the virtual test sample method. Application of the proposed classifier to the well known US Postal Service handwritten digits recognition task (USPS) yields an excellent error rate of 2.2%. We also propose a simple, but effective approach to compensate for local image transformations, which significantly increases the performance of tangent distance on a database of 1,617 medical radiographs taken from clinical daily routine.

Keywords: statistical pattern recognition, density estimation, invariant image object recognition, combined classification

1. Introduction

In this paper, a mixture density based approach to invariant image object recognition is presented. We propose a Gaussian mixture density (GMD) based Bayesian classifier and extend this non-invariant standard approach using SIMARD's tangent distance [26], as invariance plays an important role in object recognition [29]. Tangent distance is also used for the reliable estimation of covariance matrices, which is especially important if only few training samples are available. Furthermore, a new scheme for combined classification called the *virtual test sample method* (VTS) is proposed. The effectiveness of our approach is shown by applying it to the widely used US Postal Service recognition task (USPS). In the experiments, we make use of appearance based pattern recognition, i.e. each pixel of an image is interpreted as a feature, optionally performing feature reduction using a linear discriminant analysis (LDA) [10, pp. 114-123]. Using VTS and LDA, the mixture density based standard approach yields a test error rate of 3.4%. This error rate can be further improved to 2.2% by using tangent distance in the recognition step of a kernel density (KD) based classifier and by estimating the proposed *tangent covariance matrix* (without feature

reduction). To show the general applicability of the presented approach, we also apply it to a database of 1,617 medical radiograph images that were taken from the RWTH Aachen - University of Technology IRMA project (Image Retrieval in Medical Applications) [3, 18]. On this data, the performance of tangent distance can be significantly improved using a simple image distortion model to compute the proposed *distorted tangent distance.* In contrast to tangent distance, which deals with global image transformations, distorted tangent distance also compensates for local image variations.

1.1. RELATED WORK

While appearance based image object recognition is common in the pattern recognition community, the use of invariant statistical classifiers such as the one proposed here is not. MOGHADDAM & PENTLAND used Gaussian mixtures for view-based image recognition, accounting for invariances by assuming appropriate training samples and suitable image normalization [20]. SCHIELE employed histogram based image features within a Bayesian classifier, but did not use mixture densities to model the required probability densities [23]. HINTON et al. applied tangent distance to define a modified version of a principal components analysis within a linear autoencoder based classifier [14], the approach being similar to computing a maximum approximation within a mixture density based classifier. Furthermore, HASTIE et al. computed suitable prototype vectors from a given training set with respect to tangent distance, which can be used to speed up nearest neighbour classification (by using just a few prototype vectors instead of the possibly large training set) [13]. Not surprisingly, as tangent distance originated from the field of artificial neural nets, many authors such as SCHWENK use it in this context [25, 27]. An interesting review of methods for invariant pattern recognition is given in [29]. The virtual test sample method derived in Section 5 was motivated by KITTLER's research on classifier combination schemes [17]. Finally, the image distortion model used in our experiments is similar to distance measures such as the Hausdorff distance or local pertubation models. Yet, the proposed combination of this distortion model and tangent distance is a new approach.

The remainder of this paper is organized as follows: In Section 2, an overview of the databases used in the experiments is given. The GMD based standard approach is presented in Section 3, including maximum-likelihood parameter estimation, which is done by applying the Expectation-Maximization algorithm. We furthermore discuss possibilities to reduce the number of free model parameters that have to be estimated, which is crucial for successful statistical object recogni-

Figure 1. Example images taken from the original US Postal Service test set.

tion in many applications ("curse of dimensionality"). In Section 4, we introduce an affine-invariant distance measure called tangent distance (proposed by SIMARD in 1993) and its applications within the statistical classifier described here. Furthermore, we use a simple image distortion model to extend tangent distance to *distorted tangent distance*. Before presenting experimental results in Section 6, we discuss the creation of virtual data in Section 5. The virtual test sample method derived here proved to be very effective in our experiments. Finally, we will conclude the paper in Section 7.

2. Databases used in the experiments

In this section we briefly describe the image databases used in our experiments.

2.1. THE US POSTAL SERVICE DATABASE

The USPS database (available at `ftp://ftp.kyb.tuebingen.mpg.de /pub/bs/data/`) is a well known handwritten digit recognition task, which contains 7,291 training objects and 2,007 test objects. The digits are isolated and represented by a $16{\times}16$ pixels sized grayscale image (see Fig. 1). The USPS recognition task is known to be hard (commonly regarded harder as for instance the similar MNIST handwritten digits task), with a human error rate of about 2.5% on the test data [26]. An advantage of the USPS task is the availability of many recognition results reported by international research groups, allowing a fair comparison of results (cp. Tab. II). To prove that the proposed methods generalize well, we also conducted a key experiment on the MNIST handwritten digits task (60,000 reference and 10,000 test images, available at `http://www.research.att.com/~yann /ocr/mnist`).

2.2. THE IRMA DATABASE

The IRMA database consists of 110 abdomen, 706 limbs, 103 breast, 110 skull, 410 chest and 178 spine radiographs, summing up to a total

Figure 2. Example radiographs taken from the IRMA database, scaled to a common, square size. Top-left to bottom-right: abdomen, limbs, breast, skull, chest and spine.

of 1,617 images taken from daily routine (after erasing the patient information). The data is secondary digital, i.e. it has been scanned from conventional film-based radiographs. All images were scanned using 256 gray levels (with image sizes ranging from about 200×200 pixels to 2000×2000 pixels) and were labelled by an expert. Note that radiograph classification is a hard problem, as the qualities of radiographs vary considerably and there is a great within-class variance (caused by different doses of X-rays, varying orientations, pathologies or changing scribor position). Furthermore, there is a strong visual similarity between many images of the classes abdomen and spine (cp. Fig. 2). Determining the anatomic region of a given radiograph is a relevant medical problem, as this information is not available in secondary digital archives and in many cases incorrect or missing in primary digital databases. For detailed information on the motivation and the goals of the IRMA project, the reader is referred to [18].

2.3. FEATURE ANALYSIS

In our experiments we make use of *appearance based pattern recognition*, i.e. we interpret each pixel of an image as a feature. Thus, all the information contained in an image is used for classification. The only preprocessing we do for the IRMA database is downscaling the radiographs to 32×32 pixels. Our experiments showed, that this step speeds up the system significantly without notably increasing the classification error rate (cp. Section 6). Nevertheless, interpreting each pixel as a feature results in high-dimensional feature vectors (256-dimensional for USPS, 1024-dimensional for IRMA). We therefore optionally perform a linear discriminant analysis to reduce the dimensionality of the feature space, where the following calculation of the LDA transformation matrix proved to be more reliable than its straightforward calculation (the solution of a general eigenvalue problem) [10, pp. 114-123]:

In a first step, we estimate a whitening transformation matrix $\mathbf{W}$ [12, pp. 26-29]. The transformed data is called *white*, i.e. the class conditional covariance matrix

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{N} \sum_{n=1}^{N} (\boldsymbol{x}_n - \boldsymbol{\mu}_{k_n})(\boldsymbol{x}_n - \boldsymbol{\mu}_{k_n})^T \tag{1}$$

is the matrix of identity. N is the number of training samples, which are given as labelled pairs $(\mathbf{x}_n, k_n)$, $\mathbf{x}_n$ is the observation of training sample n with according class index k_n and $\boldsymbol{\mu}_{k_n}$ is the mean vector of class k_n. In a second step, we generate K prototype vectors of the form $(\boldsymbol{\mu}_k - \boldsymbol{\mu})$, where K is the number of classes, $\boldsymbol{\mu}_k$ is the mean vector of class k and $\boldsymbol{\mu}$ is the overall mean vector. These vectors are now transformed into an orthonormal basis. To avoid the numerical instabilities of the classical Gram-Schmidt approach (caused by rounding errors), this is done by using a singular value decomposition (SVD) [22, pp. 59-67], yielding a maximum of $(K-1)$ base vectors. Now, by projecting the original feature vectors into the subspace spanned, we obtain the reduced feature vectors. As the maximum number of LDA features is $(K-1)$, we define so-called *pseudoclasses* before applying the LDA to the data. These are created by performing a GMD based cluster analysis on the data (cp. Section 3). In case of the USPS database, our best LDA results were obtained creating four pseudoclasses per class, yielding 39-dimensional feature vectors.

3. The statistical approach

To classify an observation $\boldsymbol{x} \in \mathbb{R}^d$ we use the Bayesian decision rule

$$\boldsymbol{x} \longmapsto r(\boldsymbol{x}) = \underset{k}{\operatorname{argmax}} \left\{ p(k) \cdot p(\boldsymbol{x}|k) \right\}, \tag{2}$$

which is known to be optimal with respect to the expected number of classification errors in case the required distributions are known [10, pp. 10-39]. Here, $p(k)$ is the *a priori* probability of class k, $p(\boldsymbol{x}|k)$ is the *class conditional* probability for the observation $\boldsymbol{x}$ given class k and $r(\boldsymbol{x})$ is the decision of the classifier. As neither $p(k)$ nor $p(\boldsymbol{x}|k)$ are known, we have to choose models for the respective distributions and estimate their parameters using the training data. In the USPS 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), whereas on the IRMA database relative frequencies are used. The class conditional probabilities, which describe the distribution of the feature vectors in feature space, are modelled using Gaussian mixture densities

or kernel densities respectively. As the latter can be regarded as an extreme case of the mixture density model, where each training sample is interpreted as the center of a Gaussian normal distribution [3, 10, pp. 61-62], we concentrate on mixture densities in the following.

3.1. GAUSSIAN MIXTURE DENSITIES

A Gaussian mixture is defined as a linear combination of Gaussian component densities $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_{ki}, \boldsymbol{\Sigma}_{ki})$, leading to the following expression for the class conditional probabilities:

$$p(\boldsymbol{x}|k) = \sum_{i=1}^{I_k} c_{ki} \cdot \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_{ki}, \boldsymbol{\Sigma}_{ki}), \tag{3}$$

$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_{ki}, \boldsymbol{\Sigma}_{ki}) = \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}_{ki}|}} \exp\left[-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_{ki})^T \boldsymbol{\Sigma}_{ki}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_{ki})\right], \tag{4}$$

where I_k is the number of component densities used to model class k, c_{ki} are weight coefficients (with $c_{ki} > 0$ and $\sum_{i=1}^{I_k} c_{ki} = 1$), $\boldsymbol{\mu}_{ki}$ is the mean vector and $\boldsymbol{\Sigma}_{ki}$ 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, i.e. to keep the number of free model parameters small, globally pooled covariance matrices are used here:

$$\boldsymbol{\Sigma} = \sum_{k=1}^{K} \sum_{i=1}^{I_k} \frac{N_{ki}}{N} \cdot \boldsymbol{\Sigma}_{ki} \tag{5}$$

Furthermore, we only use a diagonal covariance matrix, i.e. a variance vector. Note that this does not lead to a loss of information, since a GMD of that form can still approximate any density function with arbitrary precision. Parameter estimation is now done using the Expectation-Maximization (EM) algorithm [7] in combination with a Linde-Buzo-Gray based clustering procedure [19].

3.2. PARAMETER ESTIMATION

In this section, we deal with estimating the mixture density parameters. To do so, we use the EM-algorithm, a maximum likelihood parameter estimation approach for data with so-called hidden variables. Application of the EM-algorithm to mixture densities is described in [7], where the index of some density which an observation belongs to is interpreted as hidden variable. This assignment is expressed as a probability $p(i|\boldsymbol{x}_n, k, \lambda_{ki})$, where $\lambda_{ki} = (c_{ki}, \boldsymbol{\mu}_{ki}, \boldsymbol{\Sigma}_{ki})$. By applying the

EM-algorithm, we obtain the following re-estimation formulae:

$$p(i|\boldsymbol{x}_n, k, \lambda_{ki}) \;=\; \frac{c_{ki} \cdot \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_{ki}, \boldsymbol{\Sigma}_{ki})}{\sum_{i'} c_{ki'} \cdot \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_{ki'}, \boldsymbol{\Sigma}_{ki'})} \tag{6}$$

$$\gamma_{ki}(n) \;=\; \frac{p(i|\boldsymbol{x}_n, k, \lambda_{ki})}{\sum_{n'} p(i|\boldsymbol{x}_{n'}, k, \lambda_{ki})} \tag{7}$$

$$\bar{c}_{ki} \;=\; \frac{1}{N_k} \sum_{n=1}^{N_k} p(i|\boldsymbol{x}_n, k, \lambda_{ki}) \tag{8}$$

$$\bar{\boldsymbol{\mu}}_{ki} \;=\; \sum_{n=1}^{N_k} \gamma_{ki}(n) \cdot \boldsymbol{x}_n \tag{9}$$

$$\bar{\boldsymbol{\Sigma}}_{ki} \;=\; \sum_{n=1}^{N_k} \gamma_{ki}(n) \cdot [\boldsymbol{x}_n - \bar{\boldsymbol{\mu}}_{ki}][\boldsymbol{x}_n - \bar{\boldsymbol{\mu}}_{ki}]^T \tag{10}$$

with N_k being the number of training samples of class k. The iteration is started by estimating the parameters c_{ki}, $\boldsymbol{\mu}_{ki}$ and $\boldsymbol{\Sigma}_{ki}$, yielding the initial $p(i|\boldsymbol{x}_n, k, \lambda_{ki})$. The parameters λ_{ki} are now re-estimated by setting $c_{ki} := \bar{c}_{ki}$, $\boldsymbol{\mu}_{ki} := \bar{\boldsymbol{\mu}}_{ki}$ and $\boldsymbol{\Sigma}_{ki} := \bar{\boldsymbol{\Sigma}}_{ki}$, yielding a better estimation for $p(i|\boldsymbol{x}_n, k, \lambda_{ki})$. This procedure repeateds until the parameters converge. Here, the number of densities to be trained per mixture and their initial parameters are defined by repeatedly splitting mixture components, i.e. we use a Linde-Buzo-Gray [19] inspired method. To overcome the problem of choosing the initial values for the parameters, a single density is trained for each class first. A mixture density is then created by splitting single densities, i.e. a mixture component ki is splitted by modifying the mean vector $\boldsymbol{\mu}_{ki}$ using a suitable distortion vector $\boldsymbol{\epsilon}$. In our experiments, fast convergence was obtained by choosing $\boldsymbol{\epsilon}$ to be a fraction of the respective variance vector, as this method proved to be very efficient for modelling emission probabilities in speech recognition [21]. We obtain two new mean vectors $\boldsymbol{\mu}_{ki}^{+} = \boldsymbol{\mu}_{ki} + \boldsymbol{\epsilon}$ and $\boldsymbol{\mu}_{ki}^{-} = \boldsymbol{\mu}_{ki} - \boldsymbol{\epsilon}$, i.e. a mixture density with mixture components $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_{ki}^{+}, \boldsymbol{\Sigma}_{ki})$ and $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_{ki}^{-}, \boldsymbol{\Sigma}_{ki})$. The mixture density parameters can now be re-estimated using Eq. (6)-(10), with the splitting procedure repeating until the desired number of densities is reached.

3.3. INVARIANCE PROPERTIES OF THE CLASSIFIER

Note that the appearance based statistical approach presented above is only invariant with respect to image transformations if these variabilities are present in the training data. Therefore, additional invariance of the classifier is especially useful for small training sets, as they are dealt with here. In the next sections, we therefore present two possibilities to

achieve this property, namely (a) the incorporation of invariant distance measures and (b) the generation of virtual data. In the experiments conducted, these methods proved to be superior to approaches such as the extraction of invariant features or image normalization.

4. Incorporation of invariant distance measures

We start this section by dealing with global image transformations, such as rotations or shifts. A good means to compensate for such transformations is a distance measure called *tangent distance*, which was introduced by SIMARD et al. in 1993 and which proved to be especially effective for optical character recognition [26]. In the following considerations, the class index k is dropped for ease of notation.

4.1. OVERVIEW OF TANGENT DISTANCE

In his work, SIMARD observed that reasonably small transformations of certain objects (like digits) do not affect class membership [26]. 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 (cp. Fig. 7-10). 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

$$M_x = \{t(x, \alpha) : \alpha \in \mathbb{R}^L\} \subset \mathbb{R}^{I \times J} \qquad (11)$$

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 (a visualization of this 'tangent approximation' is shown in Fig. 3):

$$\hat{M}_x = \{x + \sum_{l=1}^{L} \alpha_l \cdot T_l(x) : \alpha \in \mathbb{R}^L\} \subset \mathbb{R}^{I \times J} \qquad (12)$$

Now, the single sided tangent distance $D_T(x, \mu)$ between an image x and a reference image μ is defined as
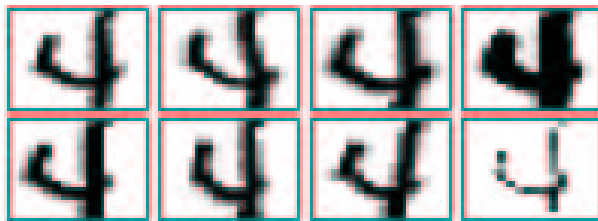
Figure 3. Example images generated via tangent approximation, using the seven tangents as proposed by Simard. The original image is at top-left.

$$D_T(\boldsymbol{x}, \boldsymbol{\mu}) = \min_{\boldsymbol{\alpha}}\{\|\boldsymbol{x} + \sum_{l=1}^{L} \alpha_l \cdot \boldsymbol{T}_l(\boldsymbol{x}) - \boldsymbol{\mu}\|^2\}. \tag{13}$$

The tangent vectors $\boldsymbol{T}_l(\boldsymbol{x})$ can be computed using simple finite differences between the original image and a small transformation of it [26]. A double sided tangent distance can also be defined by approximating $M_{\boldsymbol{x}}$ and $M_{\boldsymbol{\mu}}$ and minimizing the distance over all possible combinations of the respective parameters. In the USPS experiments, we computed the seven tangent vectors for translations (2), rotation, scaling, axis deformations (2) and line thickness, as proposed by Simard [26]. In contrast to this, the line thickness tangent loses its a-priori nature on the IRMA data and is replaced by a brightness tangent (which is set to a constant value to compensate for additive brightness variations).

Assuming that the tangent vectors are orthogonal (which can be achieved using a SVD), Eq. (13) can be solved efficiently by computing

$$D_T(\boldsymbol{x}, \boldsymbol{\mu}) = \|\boldsymbol{x} - \boldsymbol{\mu}\|^2 - \sum_{l=1}^{L} \frac{[(\boldsymbol{x} - \boldsymbol{\mu})^T \cdot \boldsymbol{T}_l(\boldsymbol{x})]^2}{\|\boldsymbol{T}_l(\boldsymbol{x})\|^2}. \tag{14}$$

The straightforward incorporation of tangent distance into the Gaussian mixture model is to replace the Mahalanobis distance by tangent distance in Eq. (4). Another approach is to use tangent approximation for reliable parameter estimation, which is treated in the following.

4.2. PARAMETER ESTIMATION WITH TANGENT APPROXIMATION

Instead of computing the empirical covariance matrix $\boldsymbol{\Sigma}$ of the given training samples $\boldsymbol{x}_1, ..., \boldsymbol{x}_N$, we can use Eq. (12) to implicitly create an "infinite" amount of training samples $\boldsymbol{x}_{n,\boldsymbol{\alpha}}$, $n = 1, ..., N$ and compute the respective *tangent covariance matrix* $\boldsymbol{\Sigma}_T$:

$$\boldsymbol{\Sigma}_T \ = \ \frac{1}{N} \int p(\boldsymbol{\alpha}) \cdot \sum_{n=1}^{N} (\boldsymbol{x}_{n,\boldsymbol{\alpha}} - \boldsymbol{\mu})(\boldsymbol{x}_{n,\boldsymbol{\alpha}} - \boldsymbol{\mu})^T \ d\boldsymbol{\alpha}, \tag{15}$$

$$x_{n,\boldsymbol{\alpha}} \;=\; x_n + \sum_{l=1}^{L} \alpha_l \cdot \boldsymbol{T}_l(x_n), \tag{16}$$

where $\boldsymbol{x}_{n,\boldsymbol{\alpha}}$ is a local transformation of the n-th training pattern, N is the number of training samples with mean $\boldsymbol{\mu}$ and $p(\boldsymbol{\alpha})$ is the distribution of the parameters $\boldsymbol{\alpha}$. With $\int p(\boldsymbol{\alpha})\, d\boldsymbol{\alpha} = 1$, $E(\boldsymbol{\alpha}) = \boldsymbol{0}$ and some elementary calculations, Eq. (15) reduces to

$$\boldsymbol{\Sigma}_T = \boldsymbol{\Sigma} + \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{T}(\boldsymbol{x}_n)\,\boldsymbol{\Sigma_{\alpha}}\,\boldsymbol{T}(\boldsymbol{x}_n)^T \tag{17}$$

with $\boldsymbol{\Sigma}$ being the empirical covariance matrix of the data, $\boldsymbol{T}(\boldsymbol{x}_n) \in \mathbb{R}^{D \times L}$ the matrix representation of the tangent vectors of training sample $\boldsymbol{x}_n$ and $\boldsymbol{\Sigma_{\alpha}} \in \mathbb{R}^{L \times L}$ the covariance matrix of the parameters $\boldsymbol{\alpha}$ (with $\boldsymbol{\Sigma_{\alpha}} = \sigma^2 \cdot \boldsymbol{I}$ in the experiments). Note that

$$\boldsymbol{\mu}_T = \frac{1}{N}\int p(\boldsymbol{\alpha}) \cdot \sum_{n=1}^{N} x_{n,\boldsymbol{\alpha}}\, d\boldsymbol{\alpha} = \boldsymbol{\mu}. \tag{18}$$

Thus, the empirical sample mean does not change in the presence of tangent vectors. More information on the probabilistic interpretation of tangent distance can be found in [4, 16].

4.3. THE IMAGE DISTORTION MODEL

Computation of tangent distance as given in Eq. (13) still requires the calculation of the (squared) Euclidean distance between the optimally transformed image $\boldsymbol{x}$ and the reference image $\boldsymbol{\mu}$. Although small global transformations have been compensated for by the optimal tangent approximation, this distance is still highly sensitive to *local* transformations of the images, e.g. caused by noise (e.g. typical for radiographs). We therefore use the following image distortion model (IDM): When calculating the distance between two images $\boldsymbol{x}$ and $\boldsymbol{\mu}$ we allow for local
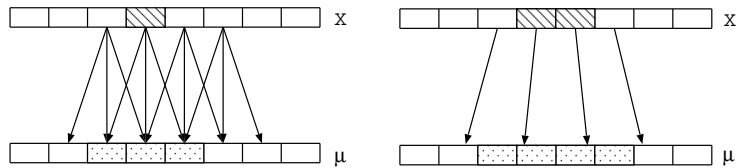

Figure 4. 1D comparison of the image distortion model (left) and the tangent model (right, showing a scale operation): The IDM allows for any locally optimal transformation, whereas the tangent model imposes global restrictions, leading to a homogeneous transformation.

deformations, i.e. we do not compute the squared error between a pixel (i, j) in $\boldsymbol{x}$ and its counterpart in $\boldsymbol{\mu}$, but look for the 'best-fitting' pixel in $\boldsymbol{\mu}$ within a certain neighbourhood R_{ij} (see Fig. 4):

$$D_{dist}(\boldsymbol{x}, \boldsymbol{\mu}) = \sum_{i=1}^{I} \sum_{j=1}^{J} \min_{(i',j') \in R_{ij}} \left\{ \|x_{ij} - \mu_{i'j'}\|^2 + C(i, i', j, j') \right\} \quad (19)$$

for images with dimension $I \times J$, where $C(i, i', j, j')$ is a cost function that models the costs for deforming a pixel x_{ij} in the input image to a pixel $\mu_{i'j'}$ in the reference image. The region R_{ij} is typically chosen to be square, resulting in a size of $(2r + 1) \times (2r + 1)$ pixels for 2D-images, with $r = 0$ yielding Euclidean distance. As the distortion distance between almost any two images can be reduced to a value near zero by increasing r (leading to a significant increase in classification error), the choice of the cost function is important for large r. In the experiments, $C(i, i', j, j')$ is chosen to be a weighted Euclidean distance between x_{ij} and $\mu_{i'j'}$. Thus, small local transformations are preferred to (most probably unwanted) long-range pixel transformations. The combination of tangent distance and the above image distortion model is called *distorted tangent distance* here and can be regarded as performing an image registration step (via optimal tangent approximation) prior to computing the image distortion distance as given in Eq. (19).

5. Virtual data creation

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.

5.1. CREATING VIRTUAL TRAINING DATA

Here, the basic idea is to choose a transformation which respects class membership and to apply it to the training samples. In the USPS experiments for example, we used ± 1 pixel shifts to create $9 \cdot 7,291 = 65,619$ training samples of size 18×18 pixels from the original 7,291 USPS training samples (of size 16×16 pixels). Thus, parameter estimation as proposed in Section 3 is not only more reliable (as there is more training data to learn from), but we also incorporate local invariances with respect to the chosen transformations into the mixture model.

5.2. 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
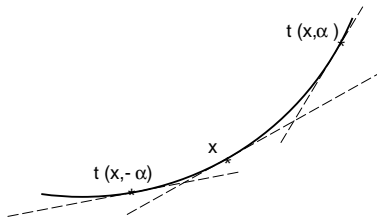
Figure 5. Approximation of the original manifold using shifted data: A 2D example.

create a number of virtual test samples $\boldsymbol{x}_1, ..., \boldsymbol{x}_{|A|}$ by applying transformations $\boldsymbol{t}(\boldsymbol{x}, \boldsymbol{\alpha})$, $\boldsymbol{\alpha} \in A$ to the given observation. On USPS, we use ± 1 pixel shifts, i.e. $|A| = 9$ for virtual data creation; other transformations such as rotation or scale might be considered in other domains, but did not improve our results. As an image cannot be shifted into different directions at the same time, the "events" $\boldsymbol{x}_1, ..., \boldsymbol{x}_{|A|}$ can be regarded as being mutually exclusive. Thus, we can model the class-conditional probability for the original observation by computing

$$
\begin{aligned}
p(\boldsymbol{x}|k) &= \sum_{\boldsymbol{\alpha} \in A} p(\boldsymbol{x}, \boldsymbol{\alpha}|k) \\
&= \sum_{\boldsymbol{\alpha} \in A} p(\boldsymbol{\alpha}) \cdot p(\boldsymbol{x}|\boldsymbol{\alpha}, k) = \frac{1}{|A|} \sum_{\boldsymbol{\alpha} \in A} p(\boldsymbol{x}_{\boldsymbol{\alpha}}|k) \quad\quad (20)
\end{aligned}
$$

(assuming equal prior probabilities $p(\boldsymbol{\alpha})$ for all transformations considered here), where the term $1/|A|$ does not depend on k and may be neglected for classification purposes. Note that this motivation for the sum rule differs from that proposed by KITTLER in [17]. Using multiple classifiers to classify a single test pattern, it was assumed that the posterior probabilities computed by the respective classifiers do not differ much from the prior probabilities in order to justify the sum rule. In contrast to this, using multiple test patterns and a single classifier, Eq. (20) 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 (computationally expensive) training phase remains unaffected. Despite its simplicity, VTS proved to be very effective in our experiments. Two things should be noted on VTS and the creation of virtual data in general:

First, creation of virtual data is not uncommon in pattern recognition. Yet, it is interesting to see that creation of virtual test samples in combination with the sum rule for combined classification is not only
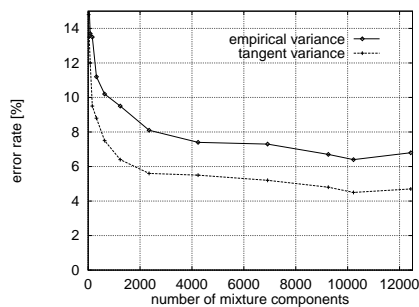
Figure 6. Empirical variance vs. tangent variance: Mixture density error rates with respect to the total number of mixture components used (9-1, no LDA).

effective, but also straightforward to justify (contrary to the multiple classifier case described in [17]). Second, combination of tangent distance and the creation of virtual data makes sense, as tangent distance is not fully invariant to - for instance - image shifts, it is only approximately invariant. Thus, creating virtual data can be interpreted as yielding a better approximation of the original manifold (cp. Fig. 5).

6. Experimental results

In this section we present some results obtained on the USPS respectively the IRMA database in our experiments.

6.1. Experiments on the usps corpus

Experiments were started by applying the GMD based standard approach to the USPS data. Table I 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 5. To compare the effectiveness of the VTS method to conventional classifier combination schemes, we used the ADABOOST algorithm [11] to boost our classifier. Doing so, the 9-1 LDA error rate dropped from 4.5% to 4.2%. Yet, by reducing the error rate from 4.5% to 3.4%, VTS significantly outperformed ADABOOST on this particular data set.

The improvements gained from the application of the LDA are mainly due to the problem of estimating variances in a high-dimensional feature space, as the next experiment shows, where we used Eq. (17) to estimate tangent variances in the EM training phase without performing feature reduction. Doing so, the error rate drops significantly from 6.0% to 4.3%. A comparison of both approaches with respect

Table I. GMD error rates on USPS with varying variance
estimation and distance measures, with and without LDA.

Method:	Error rate [%]			
	1-1	1-9	9-1	9-9
baseline	8.0	6.6	6.4	6.0
baseline + LDA	6.7	5.9	4.5	3.4
baseline + Σ_T + Mahalanobis	6.4	4.8	4.5	4.3
baseline + Σ_T + tangent	3.9	3.6	3.4	2.9

to the total number of densities used in the probabilistic model can
be found in Fig. 6. Apparently, computing tangent variances in com-
bination with the explicit creation of 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, the Mahalanobis distance used in the Gaus-
sian component densities was replaced by single sided TD in the recog-
nition step (the training phase remained unaffected), further reducing
the error rate from 4.3% to 2.9% (cp. Tab. I). This result could be fur-
ther improved to 2.7% by calculating the double sided tangent distance
in recognition (using a total of about 10,000 mixture components, i.e.
about 1,000 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 further reduced the error rate to 2.2% [2, 15].

A comparison of our USPS results with that reported by other
groups can be found in Tab. II, proving them to be excellent. Note
that results marked with an asterisk were achieved by adding about
2,400 machine printed digits to the training set [9, 26]. We also per-
formed experiments with the proposed image distortion model, Fourier
transform based invariants [6], invariant moments and discriminative
training of Gaussian mixtures [5], yet so far none of these approaches
could improve our best result of 2.2%. Furthermore, using tangent
distance in the training phase yielded no improvement.

6.2. EXPERIMENTS ON THE IRMA CORPUS

As there are only 1,617 radiographs available, a *leaving-one-out ap-
proach* was adopted here. Thus, each image was classified separately,
using the remaining 1,616 as reference images. As already mentioned in
Section 2, the radiographs were scaled down to a standard size of 32×32
pixels. This can be done without a significant change in classification
error rate, but leads to a considerable system speedup. Performing a
1-nearest neighbour classifier on the radiographs with a size of 320×320
pixels gives a classification error of 18.0%, requiring about 30 CPU

Table II. Error rates [%] reported on USPS and MNIST.

Method		USPS	MNIST
Human Performance [26]		2.5	0.2
Two-Layer Neural Net [28]		5.9	-
5-Layer Neural Net (LeNet1) [27]		4.2	1.7
Invariant Support Vectors [24]		3.0	0.8
Tangent Distance, 1-NN [26]		*2.5	1.1
Boosted Neural Net, [9]		*2.6	0.7
This work:	GMD, LDA, (VTS)	4.5 (3.4)	-
	GMD, VTS, TD	2.7	-
	KD, VTS, TD, (bagging)	2.2	1.0 (-)

seconds on a 500MHz Digital ALPHA CPU to classify a single image. Downscaling the images to the proposed size, an error rate of 18.1% was obtained, requiring about 0.4 CPU seconds per image.

We now used the single-sided tangent distance for radiograph classification. As can be seen in Tab. III, this reduces the kernel density error rate from 16.4% to 14.8% (due to the high-dimensional data, we only used class-specific standard deviations here). We then started experiments with the IDM, using $C(i, i', j, j') = 0$. Surprisingly, with an error rate of 14.7% the result of this simple distortion model is even slightly better than that obtained by using tangent distance. Computing distorted tangent distance further reduced the error rate to 12.5% (using $r = 0.7$ for the IDM), proving that the effects of the IDM and TD are additive (this could have be expected, as TD compensates for global and the IDM for local transformations).

In another experiment, the maximum local distance between two image pixels was restricted by a threshold d_{max}. Note that the maximum contribution of a pixel to any of the proposed distance measures is $255 \cdot 255 = 65,025$, as the radiographs are 256-grayscale images. Thus, a few distorted pixels (as caused by noise or changing scribor position) can cause a misclassification. By restricting this contribution to a maximum value d_{max} (called thresholding in the following), i.e. by replacing

$$\|x_{ij} - \mu_{i'j'}\|^2 \longmapsto \min\left\{\|x_{ij} - \mu_{i'j'}\|^2, d_{max}\right\} \qquad (21)$$

in Eq. (19) we can compensate for this effect, reducing the error rate to 10.3% using $d_{max} = 3500$. Analysing the remaining errors we found out that many misclassifications could be easily avoided by taking into consideration the original image aspect ratios (by downscaling the im-

Table III. Leaving-one-out IRMA error rates
[%] with respect to varying distance measures
(with and without thresholding).

Distance Measure	Thresholding	
(in kernel density)	no	yes
Mahalanobis Distance	16.4	14.2
Tangent Distance	14.8	12.9
Image Distortion Model	14.7	13.2
Distorted Tangent Distance	12.5	10.3

ages to a standard size this information is lost). To compensate for this, an aspect ratio penalty term was introduced, based on the difference in aspect ratio between the given image and the reference image. This penalty further reduced the classification error from 10.3% to 8.6%. We then chose $C(i, i', j, j')$ to be a weighted Euclidean distance between pixels (see Section 4), obtaining an error rate of 8.2%.

In a final experiment, the different distance measures discussed above were analysed with respect to their invariance properties, given a transformation t. In our experiments, we chose t to be a translation and calculated the distance between a shifted version of a radiograph and the original image as well as the distance to radiographs from competing classes. As can be seen in Fig. 7, Euclidean distance is highly sensitive to image translations. On the other hand, tangent distance (see Fig. 8) can nearly compensate one pixel shifts and yields small distances up to 2-3 pixels shifts. Naturally, the IDM with $r = 1$ (as shown in Fig. 9) can fully compensate one pixel shifts, yet with r increasing, the distances to competing classes get smaller rapidly (see Fig. 10). Thus, large neighbourhoods may lead to bad classification results.

6.3. GENERALIZATION & COMPUTATIONAL COMPLEXITY

Finally, to investigate the generalization properties of the methods presented, experiments were conducted on two completely new datasets. The best non-bagged USPS classifier was applied to the MNIST task (without doing any MNIST specific parameter optimization), whereas for the IRMA task a new dataset of 332 radiographs was collected from daily routine and then classified using the original 1,617 IRMA images as references and the parameters determined on the original IRMA data. The obtained results of 1.0% for MNIST (cp. Tab. II) and 9.0% for the new radiograph data show that the proposed methods generalize well.
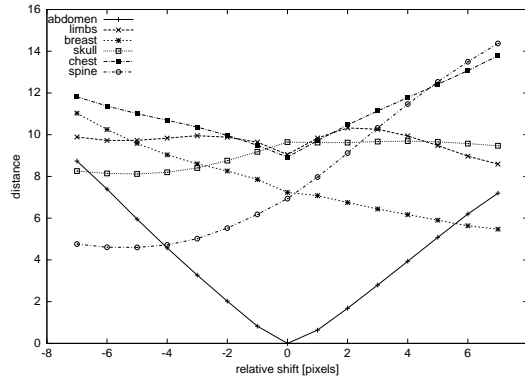
Figure 7. Behaviour of Euclidean distance with respect to image shifts.
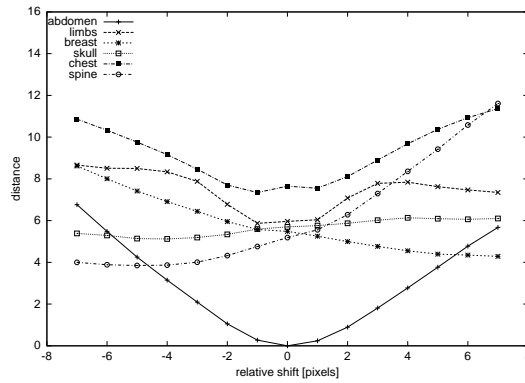

Figure 8. Behaviour of tangent distance with respect to image shifts.

Naturally, the computational complexity of the proposed method increases with the number of densities (i.e. I_k) increasing, with kernel densities being most expensive. Nevertheless, our experiments did not aim at minimizing the computational complexity, instead recognition accuracy was the most important goal. Computing single-sided/ double-sided tangent distance in a kernel density setting takes about one respectively ten seconds per image (on USPS as well as on IRMA, as the radiographs are bigger, but there are fewer reference images) on a Digital ALPHA 500 MHz CPU. Furthermore, algorithms to reduce the number of references without loosing too much classification performance are known, among them editing and condensing techniques [8].
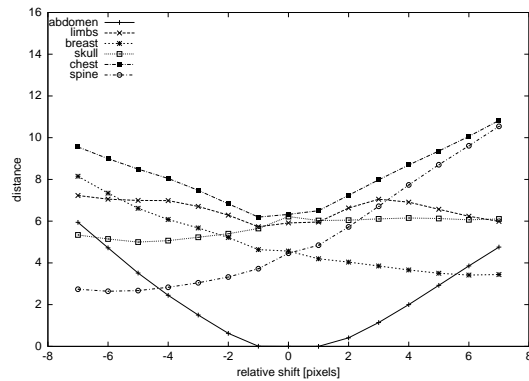
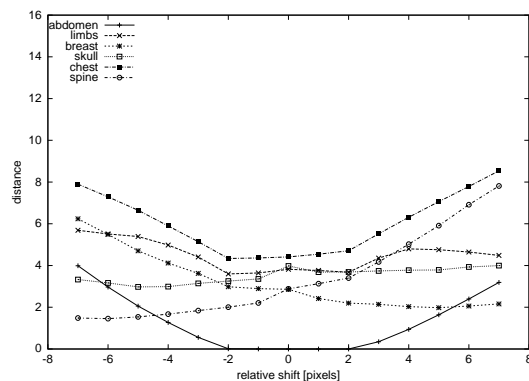*Figure 9.* Behaviour of distortion distance with respect to image shifts, using a neighbourhood with $r = 1$.

Figure 10. Behaviour of distortion distance with respect to image shifts, using a neighbourhood with $r = 2$.

7. Conclusion

In this paper, we presented an invariant, mixture density based approach to statistical image object recognition. The effectiveness of our method was shown by applying it to the well known US Postal Service handwritten digits recognition task, as well as to a completely different task, consisting of 1,617 medical radiographs. The obtained USPS error rate of 2.2% (using the original USPS training and test sets) is the best result published so far on this particular dataset. Given the difficulty of the task, the obtained error rate of 8.2% on the IRMA database of scanned radiographs is a very good result, too, proving the wide variety of possible applications of the proposed approach. On the USPS data, estimation of the proposed tangent covariance matrix proved to be especially effective, as well as using the proposed virtual test sample method. On the IRMA data it could be shown, that the image dis-

tortion model significantly reduced the tangent distance error rate by computing distorted tangent distance. As neither of the datasets used in the experiments features a development test set, the generalization abilities of the proposed methods were shown by applying the best US Postal respectively IRMA system to two completely new datasets.

Future work includes the application of the proposed algorithms in a statistical image retrieval system, where the objects detected in an image will be used as image indices. First steps towards such a system are presented in [1].

References

1. J. Dahmen, K. Beulen, M. Güld, and H. Ney, "A mixture density based approach to object recognition for image retrieval", in Proc. 6th Int. RIAO Conf. on Content-Based Multimedia Information Access, Paris, France, 2000, pp. 1632-1647.
2. J. Dahmen, D. Keysers, M. Güld, and H. Ney, "Invariant image object recognition using Gaussian mixture densities", in Proc. 15th Int. Conf. on Pattern Recognition, Barcelona, Spain, 2000, pp. 614-617.
3. J. Dahmen, T. Theiner, D. Keysers, H. Ney, T. Lehmann, and B. Wein, "Classification of radiographs in the 'Image Retrieval in Medical Applications' system (IRMA)", in Proc. 6th Int. RIAO Conf. on Content-Based Multimedia Information Access, Paris, France, 2000, pp. 551-566.
4. J. Dahmen, D. Keysers, M. Pitz, and H. Ney, "Structured covariance matrices for statistical image object recognition", in Proc. 22nd Symposium German Association for Pattern Recognition, Kiel, Germany, 2000, pp. 99-106.
5. J. Dahmen, R. Schlüter, and H. Ney, "Discriminative training of Gaussian mixtures for image object recognition", in Proc. 21st Symposium German Association for Pattern Recognition, Bonn, Germany, 1999, pp. 205-212.
6. J. Dahmen, J. Hektor, R. Perrey, and H. Ney, "Automatic classification of red blood cells using Gaussian mixture densities", in Proc. Bildverarbeitung für die Medizin, Munich, Germany, 2000, pp. 331-335.
7. A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm", Journal Royal Statistical Society, Vol.39(B), pp. 1-38, 1977.
8. P. Devijver, and J. Kittler, *Pattern Recognition. A Statistical Approach.* Prentice-Hall, Englewood Cliffs, NJ, 1982.
9. H. Drucker, R. Schapire, and P. Simard, "Boosting performance in neural networks", Int. Journal of Pattern Recognition and Artificial Intelligence, Vol. 7, No. 4, pp. 705-719, 1993.
10. R. Duda and P. Hart, *Pattern classification and scene analysis*, John Wiley & Sons, New York, NY, 1973.
11. Y. Freund and R. Schapire, "Experiments with a new boosting algorithm", in Proc. 13th Int. Conf. on Machine Learning, Bari, Italy, 1996, pp. 148-156.
12. K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press, San Diego, CA, 1990.

13. T. Hastie, P. Simard, and E. Säckinger, "Learning prototype models for tangent distance", In Proc. Advances in Neural Information Processing Systems 7, MIT Press, Cambridge, MA, 1995, pp. 999-1006.

14. G. Hinton, M. Revow, and P. Dayan, "Recognizing handwritten digits using a mixture of linear models", In Proc. Advances in Neural Information Processing Systems 7, MIT Press, Cambridge, MA, 1995, pp. 1015-1022.

15. D. Keysers, J. Dahmen, T. Theiner, and H. Ney, "Experiments with an extended tangent distance", in Proc. 15th Int. Conf. on Pattern Recognition, Barcelona, Spain, 2000, pp. 38-42.

16. D. Keysers, J. Dahmen, and H. Ney, "A Probabilistic view on tangent distance", in Proc. 22nd Symposium German Association for Pattern Recognition, Kiel, Germany, 2000, pp. 107-114.

17. J. Kittler, M. Hatef, and R. Duin, "Combining classifiers", in Proc. 13th Int. Conf. on Pattern Recognition, Vienna, Austria, 1996, pp. 897-901.

18. T. Lehmann, B. Wein, J. Dahmen, J. Bredno, F. Vogelsang, and M. Kohnen, "Content-based image retrieval in medical applications: a novel multi-step approach", in Proc. Int. Society for Optical Engineering (SPIE), Vol. 3972, No. 32, 2000, pp. 312-320.

19. Y. Linde, A. Buzo, and R. Gray, "An algorithm for vector quantizer design", IEEE Trans. Communications, Vol. 28, No. 1, pp. 84-95, 1980.

20. B. Moghaddam and A. Pentland, "Probabilistic visual learning for object representation", IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 19, No. 7, pp. 696-710, July 1997.

21. H. Ney, "Acoustic modelling of phoneme units for continuous speech recognition", L. Torres, E. Masgrau, M. Lagunas (eds.): *Signal Processing V: Theories and Applications*, Elsevier Science Publishers B.V., 1990.

22. W. Press, S. Teukolsky, W. Vetterling, and B. Flannery, *Numerical Recipes in C*, University Press, Cambridge, NY, 1992.

23. B. Schiele and J. Crowley, "Probabilistic object recognition using multidimensional receptive field histograms", in Proc. 13th Int. Conf. on Pattern Recognition, Vienna, Austria, 1996, pp. 50-54.

24. B. Schölkopf, P. Simard, A. Smola, and V. Vapnik, "Prior knowledge in support vector kernels", In Proc. Advances in Neural Information Processing Systems 10, MIT Press, Cambridge, MA, 1998, pp. 640-646.

25. H. Schwenk and M. Milgram, "Transformation invariant autoassociation with application to handwritten character recognition", In Proc. Advances in Neural Information Processing Systems 7, MIT Press, Cambridge, MA, 1995, pp. 991-998.

26. P. Simard, Y. Le Cun, and J. Denker, "Efficient pattern recognition using a new transformation distance", In Proc. Advances in Neural Information Processing Systems 5, Morgan Kaufmann, San Mateo, CA, pp. 50-58, 1993.

27. 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, New York, NY, pp. 239-274, 1998.

28. V. Vapnik. *The nature of statistical learning theory*. Springer, New York, NY, 1995.

29. J. Wood, "Invariant Pattern Recognition: A Review", Pattern Recognition, Vol. 29, No. 1, pp. 1–17, January 1996.