Learning of Variability for Invariant Statistical Pattern Recognition

Daniel Keysers, Wolfgang Macherey, Jörg Dahmen, and Hermann Ney

Lehrstuhl für Informatik VI, Computer Science Department
RWTH Aachen - University of Technology, D-52056 Aachen, Germany
{keysers, w.macherey, dahmen, ney}@informatik.rwth-aachen.de
WWW home page: http://www-i6.informatik.rwth-aachen.de

Abstract. In many applications, modelling techniques are necessary which take into account the inherent variability of given data. In this paper, we present an approach to model class specific pattern variation based on tangent distance within a statistical framework for classification. The model is an effective means to explicitly incorporate invariance with respect to transformations that do not change class-membership like e.g. small affine transformations in the case of image objects. If no prior knowledge about the type of variability is available, it is desirable to learn the model parameters from the data. The probabilistic interpretation presented here allows us to view learning of the variational derivatives in terms of a maximum likelihood estimation problem. We present experimental results from two different real-world pattern recognition tasks, namely image object recognition and automatic speech recognition. On the US Postal Service handwritten digit recognition task, learning of variability achieves results well comparable to those obtained using specific domain knowledge. On the SieTill corpus for continuously spoken telephone line recorded German digit strings the method shows a significant improvement in comparison with a common mixture density approach using a comparable amount of parameters. The probabilistic model is well-suited to be used in the field of statistical pattern recognition and can be extended to other domains like cluster analysis.

1 Introduction

In many applications, it is important to carefully consider the inherent variability of data. In the field of pattern recognition it is desired to construct classification algorithms which tolerate variation of the input patterns that leaves the class-membership unchanged. For example, image objects are usually subject to affine transformations of the image grid like rotation, scaling and translation. Conventional distance measures like the Euclidean distance or the Mahalanobis distance [3] do not take into account such transformations or do so only if the training data contains a large number of transformed patterns, respectively. One method to incorporate *invariance* whith respect to such transformations into a classifier is to use invariant distance measures like the *tangent distance*, which has been successfully applied in image object recognition during the last years [9, 14, 15].

Tangent distance (TD) is usually applied by explicitly modelling the derivative of transformations which are known a priori. This is especially effective in cases where the training set is small. But not in all domains such specific knowledge is available. For example, the transformation effects on the feature vectors of a speech signal that are used in automatic speech recognition are generally difficult to obtain or unknown.

In this paper we present a method to automatically *learn* the derivative of the variability present in the data within a statistical framework, thus leading to an increased robustness of the classifier. To show the practical value of the approach we present results from experiments in two real-world application areas, namely optical character recognition (OCR) and automatic speech recognition (ASR).

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

$$x \longmapsto r(x) = \operatorname*{argmax}_k \{p(k) \cdot p(x|k)\} . \tag{1}$$

Here, $p(k)$ 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 decision of the classifier. This decision rule is known to be optimal with respect to the expected number of classification errors if the required distributions are known [3]. However, as neither $p(k)$ nor $p(x|k)$ are known in practical situations, it is necessary to choose models for the respective distributions and estimate their parameters using the training data. The class conditional probabilities are modelled using *Gaussian mixture densities* (GMD) or *kernel densities* (KD) in the experiments. 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 distribution. A Gaussian mixture is defined as a linear combination of Gaussian component densities, which can approximate any density function with arbitrary precision, even if only component densities with diagonal covariance matrices are used. This restriction is often imposed in order to reduce the number of parameters that must be estimated. The necessary parameters for the GMD can be estimated using the Expectation-Maximization (EM) algorithm [3].

2 Invariance and Tangent Distance

There exists a variety of ways to achieve invariance or transformation tolerance of a classifier, including normalization, extraction of invariant features and invariant distance measures [19]. Distance measures are used for classification as dissimilarity measures, i.e. the distances should ideally be small for members of the same class and large for members of different classes. An invariant distance measure ideally takes into account transformations of the patterns, yielding small values for patterns which mostly differ by a transformation that does not change class-membership. In the following, we will give a brief overview of one invariant distance measure called *tangent distance*, which was introduced in [15, 16].

Let $x \in \mathbb{R}^D$ be a pattern and $t(x, \alpha)$ denote a transformation of x that depends on a parameter L-tuple $\alpha \in \mathbb{R}^L$, where we assume that t does not

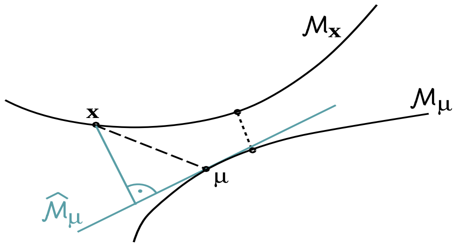

Fig. 1. Illustration of the Euclidean distance between an observation x and a reference μ (dashed line) in comparison to the distance between the corresponding manifolds (dotted line). The tangent approximation of the manifold of the reference and the corresponding (one-sided) tangent distance is depicted by the light gray lines.

affect class membership (for small α). The set of all transformed patterns now comprises a manifold $\mathcal{M}_x = \{t(x, \alpha) \ : \ \alpha \in \mathbb{R}^L\} \subset \mathbb{R}^D$ in pattern space. The distance between two patterns can then be defined as the minimum distance between the manifold $\mathcal{M}_x$ of the pattern x and the manifold $\mathcal{M}_\mu$ of a class specific prototype pattern μ, which is truly invariant with respect to the regarded transformations (cf. Fig. 1):

$$d(x, \mu) = \min_{\alpha, \beta \in \mathbb{R}^L} \{||t(x, \alpha) - t(\mu, \beta)||^2\} \tag{2}$$

However, the resulting distance calculation between manifolds is a hard non-linear optimization problem in general. Moreover, the manifolds usually cannot be handled analytically. To overcome these problems, the manifolds can be approximated by a *tangent subspace* $\widehat{\mathcal{M}}$. The *tangent vectors* x_l that span the subspace are the partial *derivatives* of the transformation t with respect to the parameters α_l ($l = 1, \ldots, L$), i.e. $x_l = \partial t(x, \alpha)/\partial \alpha_l$. Thus, the transformation $t(x, \alpha)$ can be approximated using a Taylor expansion around $\alpha = 0$:

$$t(x, \alpha) = x + \sum_{l=1}^{L} \alpha_l x_l + \sum_{l=1}^{L} \mathcal{O}(\alpha_l^2) \tag{3}$$

The set of points consisting of all linear combinations of the pattern x with the tangent vectors x_l forms the tangent subspace $\widehat{\mathcal{M}}_x$, which is a first-order approximation of $\mathcal{M}_x$:

$$\widehat{\mathcal{M}}_x = \{x + \sum_{l=1}^{L} \alpha_l x_l \ : \ \alpha \in \mathbb{R}^L\} \subset \mathbb{R}^D \tag{4}$$

Using the linear approximation $\widehat{\mathcal{M}}_x$ has the advantage that distance calculations are equivalent to the solution of linear least square problems or equivalently

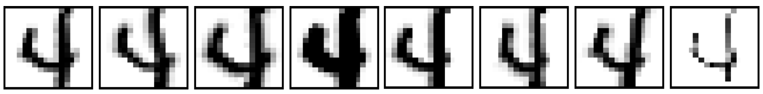

Fig. 2. Example of first-order approximation of affine transformations and line thickness. (Left to right: original image, diagonal deformation, scale, line thickness increase, shift left, axis deformation, line thickness decrease)

projections into subspaces, which are computationally inexpensive operations. The approximation is valid for small values of α, which nevertheless is sufficient in many applications, as Fig. 2 shows for examples of OCR data. These examples illustrate the advantage of TD over other distance measures, as the depicted patterns all lie in the same subspace. The TD between the original image and any of the transformations is therefore zero, while the Euclidean distance is significantly greater than zero. Using the squared Euclidean norm, the TD is defined as:

$$d_{2S}(x,\mu) = \min_{\alpha,\beta\in\mathbb{R}^L} \{ ||(x + \sum_{l=1}^{L} \alpha_l x_l) - (\mu + \sum_{l=1}^{L} \beta_l \mu_l)||^2 \} \tag{5}$$

Eq. (5) is also known as *two-sided* tangent distance (2S) [3]. In order to reduce the effort for determining $d_{2S}(x,\mu)$ it may be convenient to restrict the calculation of the tangent subspaces to the prototype (or the reference) vectors. The resulting distance measure is called *one-sided* tangent distance (1S):

$$d_{1S}(x,\mu) = \min_{\alpha\in\mathbb{R}^L} \{ ||x - (\mu + \sum_{l=1}^{L} \alpha_l \mu_l)||^2 \} \tag{6}$$

The presented considerations are based on the Euclidean distance, but equally apply when using the Mahalanobis distance [3] in a statistical framework. They show that a suitable first-order model of variability is a subspace model based on the derivatives of transformations that respect class-membership, where the variation is modelled by the tangent vectors or subspace components, respectively. In the following we will concentrate on properties of the model and the estimation of subspace components if the transformations are not known.

3 Learning of Variability

We first discuss a probabilistic framework for TD and then show, how learning of the tangent vectors can be considered as the solution of a maximum likelihood estimation problem. This estimation is especially useful for cases where no prior knowledge about the transformations present in the data is available.

3.1 Tangent Distance in a Probabilistic Framework

To embed the TD into a statistical framework we will focus on the one-sided TD, assuming that the references are subject to variations. A more detailed presentation including the remaining cases of variation of the observations and the two-sided TD can be found in [8].

We restrict our considerations here to the case where the observations x are normally distributed with expectation μ and covariance matrix Σ. The extension to Gaussian mixtures or kernel densities is straightforward using maximum approximation or the EM algorithm. In order to simplify the notation, class indices are omitted. Using the first-order approximation of the manifold $\mathcal{M}_\mu$ for a mean vector μ, we obtain the probability density function (pdf) for the observations:

$$p(x \mid \mu, \alpha, \Sigma) = \mathcal{N}(x \mid \mu + \sum_{l=1}^{L} \alpha_l \mu_l,\, \Sigma) \tag{7}$$

The integral of the joint distribution $p(x, \alpha \,|\, \mu, \Sigma)$ over the unknown transformation parameters α then leads to the following distribution:

$$\begin{aligned} p(x \,|\, \mu, \Sigma) &= \int p(x, \alpha \,|\, \mu, \Sigma) \; d\alpha \\ &= \int p(\alpha \,|\, \mu, \Sigma) \cdot p(x \,|\, \mu, \alpha, \Sigma) \; d\alpha \\ &= \int p(\alpha) \cdot p(x \,|\, \mu, \alpha, \Sigma) \; d\alpha \end{aligned} \tag{8}$$

Without loss of generality, the tangent vectors of the pdf in Eq. (7) can be assumed orthonormal with respect to Σ, as only the spanned subspace determines the modelled variation. Hence, it is always possible to achieve the condition

$$\mu_l^T \Sigma^{-1} \mu_m = \delta_{l,m} \tag{9}$$

using e.g. a singular value decomposition, where $\delta_{l,m}$ denotes the Kronecker delta. Note that we assume that α is independent of μ and Σ, i.e. $p(\alpha \,|\, \mu, \Sigma) \equiv p(\alpha)$. Furthermore, $\alpha \in \mathbb{R}^L$ is assumed to be normally distributed with mean 0 and a covariance matrix $\gamma^2 I$, i.e. $p(\alpha) = \mathcal{N}(\alpha \,|\, 0, \gamma^2 I)$, where I denotes the identity matrix and γ is a hyperparameter describing the standard deviation of the transformation parameters. These assumptions reduce the complexity of the calculations but do not affect the general result. The evaluation of the integral in Eq. (8) leads to the following expression:

$$p(x|\mu, \Sigma) = \mathcal{N}(x|\mu, \Sigma') = \det(2\pi\Sigma')^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\Big[(x-\mu)^T \Sigma'^{-1} (x-\mu)\Big]\right) \tag{10}$$

$$\Sigma' = \Sigma + \gamma^2 \sum_{l=1}^{L} \mu_l \mu_l^T, \qquad \Sigma'^{-1} = \Sigma^{-1} - \frac{1}{1+\frac{1}{\gamma^2}} \Sigma^{-1} \sum_{l=1}^{L} \mu_l \mu_l^T \Sigma^{-1} \tag{11}$$

Note that the exponent in Eq. (10) leads to the conventional Mahalanobis distance for $\gamma \to 0$ and to TD for $\gamma \to \infty$. Thus, the incorporation of tangent vectors adds a corrective term to the Mahalanobis distance that only affects the covariance matrix which can be interpreted as structuring Σ [8]. For the limiting case $\Sigma = I$, a similar result was derived in [6]. The probabilistic interpretation of TD can also be used for a more reliable estimation of the parameters of the distribution [2, 8]. Note furthermore that $\det(\Sigma') = (1+\gamma^2)^L \det(\Sigma)$ [5, pp. 38ff.] which is independent of the tangent vectors and can therefore be neglected in the following maximum likelihood estimation.

3.2 Estimation of Subspace Components

In order to circumvent the restriction that the applicable transformations must be known a priori, the tangent vectors can be learned from the training data. This estimation can be formulated within a maximum likelihood approach.

Let the training data be given by $x_{n,k}, n = 1, \ldots, N_k$ training patterns of $k = 1, \ldots, K$ classes. Assuming that the number L of tangent vectors is known (note that L can be determined automatically [1]) we consider the log-likelihood as a function of the unknown tangent vectors $\{\mu_{kl}\}$ (for each class k):

$$\begin{aligned} F(\{\mu_{kl}\}) &:= \sum_{k=1}^{K}\sum_{n=1}^{N_k} \log \mathcal{N}(x_{n,k}|\mu_k, \Sigma'_k) \\ &= \frac{1}{1+\frac{1}{\gamma^2}} \sum_{k=1}^{K}\sum_{n=1}^{N_k}\sum_{l=1}^{L} ((x_{n,k}-\mu_k)^T \Sigma^{-1}\mu_{kl})^2 + \text{const} \\ &= \frac{1}{1+\frac{1}{\gamma^2}} \sum_{k=1}^{K}\sum_{l=1}^{L} \mu_{kl}^T \Sigma^{-1} S_k \Sigma^{-1}\mu_{kl} + \text{const} \end{aligned} \tag{12}$$

with $S_k = \sum_{n=1}^{N_k}(x_{n,k}-\mu_k)(x_{n,k}-\mu_k)^T$ as the class specific scatter matrix. Σ and S_k can be regarded as covariance matrices of two competing models. Taking the constraints of orthonormality of the tangent vectors with respect to Σ^{-1} into account, we obtain the following result [5, pp. 400ff.]: The class specific tangent vectors μ_{kl} maximizing Eq. (12) have to be chosen such that the vectors $\Sigma^{-1/2}\mu_{kl}$ are those eigenvectors of the matrix $\Sigma^{-1/2}S_k(\Sigma^{-1/2})^T$ with the largest corresponding eigenvalues.

As the above considerations show, two different models have to be determined for the covariance matrices Σ and S_k. While S_k is defined as a class specific scatter matrix, a globally pooled covariance matrix is a suitable choice for Σ in many cases. Using these models, the effect of incorporating the tangent distance into the Mahalanobis distance is equivalent to performing a global whitening transformation of the feature space and then using the L class specific eigenvectors with the largest eigenvalues as tangent vectors for each class. This reduces the effect of those directions of class specific variability that contribute the most variance to Σ. While the maximum likelihood estimate leads to results similar to conventional principal component analysis (PCA), the estimated components are used in a completely different manner here. In conventional PCA, the principal components are chosen to minimize the reconstruction error. In contrast to that, these components span the subspace with minor importance in the distance calculation in the approach presented here. This can be interpreted as reducing the effect of specific variability, motivated by the fact that it does not change class membership of the patterns. The tangent distance has the property that it also works very well in combination with global feature transformations as for instance a linear discriminant analysis (LDA), since Σ can be assumed as a global covariance matrix of an LDA-transformed feature space.

4 Experimental Results

To show the applicability of the proposed learning approach, we present results obtained on two real-world classification tasks. The performance of a classifier is measured by the obtained *error rate* (ER), i.e. the ratio of misclassifications to the total number of classifications. For speech recognition a suitable measure is the *word error rate* (WER), which is defined as the ratio of the number of incorrectly recognized words to the total number of words to be recognized. The difference to the correct sentence is measured using the Levenshtein or edit distance, defined as the minimal number of insertions (ins), deletions (del) or replacements of words necessary to transform the correct sentence to the recognized sentence. The *sentence error rate* (SER) is defined as the fraction of incorrectly recognized sentences.

4.1 Image Object Recognition

Results for the domain of image object recognition were obtained on the well known US Postal Service handwritten digit recognition task (USPS). It contains normalized greyscale images of size 16×16 pixels, divided into a training set of 7,291 images and a test set of 2,007 images. Reported recognition error rates for this database are summarized in Table 1. In our preliminary experiments, we used kernel densities to model the distributions in Bayes' decision rule and we applied *appearance based* classification, i.e. no feature extraction was applied. The use of tangent distance based on derivatives (6 affine derivatives plus line thickness) and virtual training and testing data (by shifting the images 1 pixel into 8 directions, keeping training and test set separated) improved the error rate to 2.4%. This shows the effectivity of the tangent distance approach in combination with prior knowledge. Finally, using classifier combination, where different test results were combined using the sum rule, we obtained an error rate of 2.2% [9].

For our experiments on learning of variability, we used two different settings. First, we used a single Gaussian density, i.e. one reference per class, and varied the number of estimated tangents. As shown in Table 2, the error rate can

Table 1. Summary of results for the USPS corpus (error rates, [%]).
*: training set extended with 2,400 machine-printed digits

method			ER[%]
human performance	[Simard et al. 1993] [15]		2.5
relevance vector machine	[Tipping et al. 2000] [17]		5.1
neural net (LeNet1)	[LeCun et al. 1990] [14]		4.2
invariant support vectors	[Schölkopf et al. 1998] [13]		3.0
neural net + boosting	[Drucker et al. 1993] [14]		*2.6
tangent distance	[Simard et al. 1993] [15]		*2.5
nearest neighbor classifier	[9]		5.6
mixture densities	[2] baseline		7.2
	+ LDA + virtual data		3.4
kernel densities	[9] tangent distance, derivative,	one-sided (μ)	3.7
		one-sided (x)	3.3
		two-sided	3.0
	+ virtual data		2.4
	+ classifier combination		**2.2**
kernel densities	tangent distance, learned, one-sided (μ), $L = 12$		3.7

Table 2. Results for learning of tangent vectors (ER [%], USPS, KD)

#references/class	$L = 0$	$L = 7$	$L = 12$	$L = 20$	derivative tangent vectors ($L = 7$)
1	18.6	6.4	5.5	5.5	11.8
≈700	5.5	3.8	3.9	3.7	3.7

be reduced from 18.6% to 5.5% with the estimation of tangent vectors from class specific covariance matrices as proposed above. Using only $L = 7$ tangent vectors, the result of 6.4% compares favorably to the use of the derivative, here with 11.8% error rate. This is probably due to the fact that the means of the single densities are the average of a large number of images and therefore very blurred, which is a disadvantage for the derivative tangent vectors. Here, the estimated tangent vectors outperform those based on the derivative.

Interestingly, when using all 7,291 training patterns in a kernel density based classifier, the result obtained without tangent model is the same as for a single density model with 12 estimated tangents. In this case, the single densities with estimated tangent subspace obtain the same result using about 50 times fewer parameters. In the second setting with about 700 references per class (KD), the error rate can be reduced to 3.7% for 20 estimated tangents. Fig. 3(a) shows the evolution of the error rate for different number of tangent vectors. Here, the tangent vectors were estimated using a local, class specific covariance matrix obtained from the set of local nearest neighbors for each training pattern. Therefore, the method is only applied to the one-sided tangent distance with tangents on the side of the reference. The obtained error rate is the same as for the derivative tangents, although somewhat higher for the same number of tangents. This shows that the presented method can be effectively used to learn the class specific variability on this dataset. Note that using the tangents on the side of the observations resp. on both sides, the obtained error rate is significantly lower (cf. Table 1).

Fig. 3(b) shows the error rate with respect to the subspace standard deviation γ for derivative tangents and estimated tangents using $L = 7$ each. It can be seen that, on this data, no significant improvement can be obtained by restricting the value of γ, while there may be improvements for other pattern recognition tasks.

So far we have not discussed the computational complexity of the tangent method. Due to the structure of the resulting model, the computational cost of the distance calculation is increased approximately by a factor of $(L + 1)$, in comparison with the baseline model that corresponds to the Euclidean distance.

4.2 Automatic Speech Recognition

Experiments for the domain of speech recognition were performed on the *SieTill* corpus [4] for telephone line recorded German continuous digit strings. The corpus consists of approximately 43k spoken digits in 13k sentences for both training and test set. In Table 3 some information on corpus statistics is summarized.

The recognition system is based on whole-word Hidden Markov Models (HMMs) using continuous emission densities. The baseline system is characterized as follows:

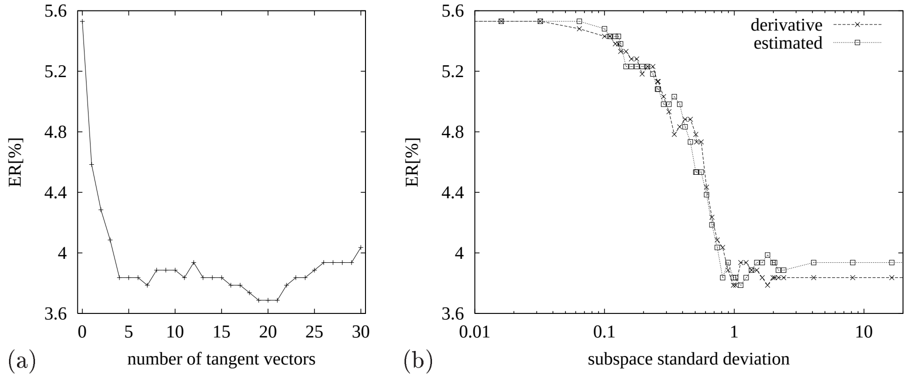

Fig. 3. (a) ER w.r.t. number of estimated tangents (USPS, KD). (b) ER w.r.t. subspace standard deviation γ for $L = 7$ derivative and estimated tangent vectors (USPS, KD).

– vocabulary of 11 German digits including the pronunciation variant '*zwo*',
– gender-dependent whole-word HMMs, with every two subsequent states being identical,
– for each gender 214 distinct states plus one for silence,
– Gaussian mixture emission distributions,
– one globally pooled diagonal covariance matrix Σ,
– 12 cepstral features plus first derivatives and the second derivative of the first feature component.

The baseline recognizer applies maximum likelihood training using the Viterbi approximation in combination with an optional LDA. A detailed description of the baseline system can be found in [18]. The word error rates obtained with the baseline system for the combined recognition of both genders are summarized in Table 4 (in the lines with 0 tangent vectors (tv) per mixture (mix)). In this domain, all densities of the mixtures for the states of the HMMs are regarded as separate *classes* for the application of learning of variability. The S_k were trained as state specific full covariance matrices. Note that the S_k are only necessary in the training phase.

For single densities, the incorporation of TD improved the word error rate by 18.1% relative for one tangent vector and 21.6% relative using four tangent vectors per state. In combination with LDA transformed features the relative improvement was 13.8% for the incorporation of one tangent vector and increased to 28.6% for five tangent vectors per state. Fig. 4(a) depicts the evolution of the word error rates on the *SieTill* test corpus for different numbers of tangent vectors using single densities that were trained on LDA transformed features.

Table 3. Corpus statistics for the SieTill corpus.

corpus	female		male	
	sent.	digits	sent.	digits
test	6176	20205	6938	22881
train	6113	20115	6835	22463

For this setting the optimal choice for gender dependent trained references was five tangent vectors per state.

Using mixture densities, the performance gain in word error rate decreased but was still significant. Thus the relative improvement between the baseline result and tangent distance was 6.7% (16 densities plus one tangent vector per mixture) for untransformed features and 13.6% for LDA transformed features (16 dns/mix, 1 tv/mix). The same applies for the optimal number of tangent vectors which was found at one tangent vector per mixture. Consequently, a larger number of densities is able to partially compensate for the error that is made in the case that the covariance matrix is estimated using the conventional method. The best result was obtained using 128 densities per mixture in combination with LDA transformed features and the incorporation of one tangent vector per state. Using this setting, the word error rate decreased from 1.85% to 1.67% which is a relative improvement of 5%. Fig. 4(b) depicts the evolution of word error rates for conventional training in comparison with TD using equal numbers of parameters. Even though the incorporation of tangent vectors into the Mahalanobis distance increases the number of parameters, the overall gain in performance justifies the higher expense.

5 Discussion and Conclusion

In this paper we presented an approach for modelling and learning variability for statistical pattern recognition, embedding tangent distance into a probabilistic framework. In contrast to principal component analysis based methods like [12] the model disregards the specific variability of the patterns when determining the distance or the log-likelihood, respectively, which leads to an incorporation of transformation tolerance and therefore improves the classification performance. This is due to the basic difference between the *distance in feature space* and the

Table 4. Word error rates (WER) and sentence error rates (SER) on the SieTill corpus obtained with the tangent distance. In column 'tv/mix' the number of used tangent vectors per mixture is given. A value of 0 means that the conventional Mahalanobis distance is used. 'dns/mix' gives the average number of densities per mixture.

without LDA					with LDA				
dns/mix	tv/mix	error rates [%]			dns/mix	tv/mix	error rates [%]		
		del - ins	WER	SER			del - ins	WER	SER
1	0	1.17-0.83	4.59	11.34	1	0	0.71-0.63	3.78	9.74
	1	1.17-0.52	3.76	9.22		1	0.97-0.49	3.26	8.46
	4	0.69-1.07	3.60	9.10		**5**	**0.48-0.88**	**2.70**	**7.18**
16	0	0.59-0.83	2.67	6.92	16	0	0.44-0.68	2.28	5.92
	1	0.54-0.58	2.49	6.56		1	0.58-0.40	1.97	5.06
	4	0.46-0.80	2.60	6.76		4	0.38-0.55	1.97	5.35
128	0	0.52-0.54	2.24	5.87	128	0	0.45-0.39	1.85	4.94
	1	0.50-0.48	2.12	5.75		**1**	**0.42-0.34**	**1.67**	**4.50**
	4	0.55-0.49	2.13	5.71		4	0.39-0.41	1.76	4.81

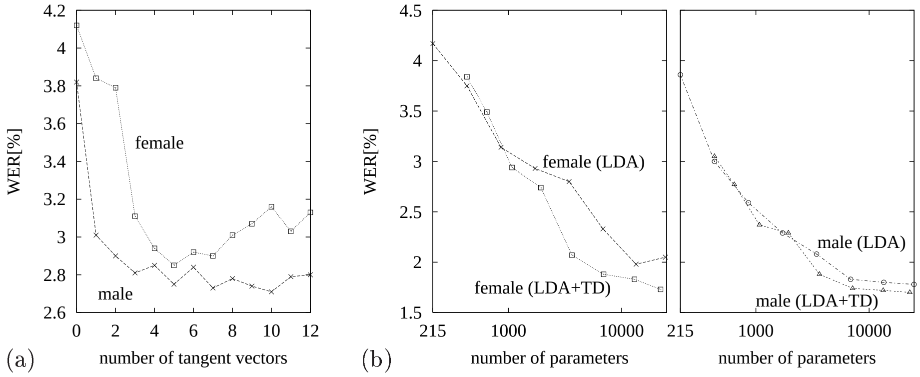

Fig. 4. (a) Word error rates as a function of the number of tangent vectors on the SieTill test corpus for single densities using ML training on LDA transformed features. (b) Comparison of WER for mixture densities on the SieTill test corpus using equal overall model parameter numbers.

distance from feature space, which seems to be more appropriate for classification [11]. The presented model in its local version is adaptive to specific local variability and therefore similar to [7]. Note that the presented model assigns to the subspace components a weight γ that was found to be usually larger than the corresponding eigenvalue, which is a main difference to subspace approximations to the full covariance matrix based on eigenvalue decomposition like e.g. [10]. The overrepresentation of estimated variational subspace components may lead to an increased transformation tolerance. The new model proved to be very effective for pattern recognition, including the combination with globally operating feature transformations as the linear discriminant analysis. Thus, theoretical findings are supported by the experimental results. Comparative experiments were performed on the USPS corpus for image object recognition and on the *SieTill* corpus for continuous German digit strings for automatic speech recognition. On the USPS corpus, single density and kernel density error rates could be significantly improved, and the obtained results were well comparable to the use of tangents based on prior knowledge. Using the one-sided TD, a relative improvement in word error rate of approximately 20% was achieved for single densities on the *SieTill* corpus. For mixture densities we could gain a relative improvement of up to 13.6% in word error rate. Incorporating the TD we were able to reduce the word error rate of our best recognition result based on maximum likelihood trained references from 1.85% to 1.67%. Note that the probabilistic modelling technique may also be used for other tasks like clustering, where first results show that the formed clusters respect the transformations.

References

1. C. M. Bishop. Bayesian PCA. In M. Kearns, S. Solla, and D. Cohn, editors, *Advances in Neural Information Processing Systems 11*. MIT Press, pages 332–388, 1999.
2. J. Dahmen, D. Keysers, H. Ney, and M. O. Güld. Statistical Image Object Recognition using Mixture Densities. *Journal of Mathematical Imaging and Vision*, 14(3):285–296, May 2001.

3. R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. John Wiley & Sons, Inc., New York, 2nd edition, 2000.
4. T. Eisele, R. Haeb-Umbach, and D. Langmann. A comparative study of linear feature transformation techniques for automatic speech recognition. In *Proc. of Int. Conf. on Spoken Language Processing*, volume I, Philadelphia, PA, pages 252–255, Oct. 1996.
5. K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Computer Science and Scientific Computing Academic Press Inc., San Diego, CA, 2nd edition, 1990.
6. T. Hastie and P. Simard. Metrics and Models for Handwritten Character Recognition. *Statistical Science*, 13(1):54–65, January 1998.
7. T. Hastie and R. Tibshirani. Discriminative Adaptive Nearest Neighbor Classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(6):607–616, June 1996.
8. D. Keysers, J. Dahmen, and H. Ney. A Probabilistic View on Tangent Distance. In *22. DAGM Symposium Mustererkennung 2000*, Springer, Kiel, Germany, pages 107–114, September 2000.
9. D. Keysers, J. Dahmen, T. Theiner, and H. Ney. Experiments with an Extended Tangent Distance. In *Proceedings 15th International Conference on Pattern Recognition*, volume 2, Barcelona, Spain, pages 38–42, September 2000.
10. P. Meinicke and H. Ritter. Local PCA Learning with Resolution-Dependent Mixtures of Gaussians. In *Proc. of ICANN'99, 9th Intl. Conf. on Artificial Neural Networks, Edinburgh, UK*, pages 497–502, September 1999.
11. B. Moghaddam and A. Pentland. Probabilistic Visual Learning for Object Representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):696–710, July 1997.
12. T. R. Payne and P. Edwards. Dimensionality Reduction through Sub-space Mapping for Nearest Neighbor Algorithms. In *Proceedings ECML 2000, 11th European Conference on Machine Learning*, volume 1810 of *Lecture Notes in Artificial Intelligence*, Springer, Barcelona, Spain, pages 331–343, May 2000.
13. B. Schölkopf, P. Simard, A. Smola, and V. Vapnik. Prior Knowledge in Support Vector Kernels. In M. I. Jordan, M. J. Kearns, and S. A. Solla, editors, *Advances in Neural Inf. Proc. Systems*, volume 10. MIT Press, pages 640–646, 1998.
14. P. Simard, Y. Le Cun, J. Denker, and B. Victorri. Transformation Invariance in Pattern Recognition — Tangent Distance and Tangent Propagation. In G. Orr and K.-R. Müller, editors, *Neural networks: tricks of the trade*, volume 1524 of *Lecture Notes in Computer Science*, Springer, Heidelberg, pages 239–274, 1998.
15. 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 Inf. Proc. Systems*, volume 5, Morgan Kaufmann, San Mateo CA, pages 50–58, 1993.
16. P. Simard, Y. Le Cun, J. Denker, and B. Victorri. An Efficient Algorithm for Learning Invariances in Adaptive Classifiers. In *Proceedings 11th International Conference on Pattern Recognition*, The Hague, The Netherlands, pages 651–655, August 1992.
17. M. E. Tipping. The Relevance Vector Machine. In S. Solla, T. Leen, and K. Müller, editors, *Advances in Neural Information Processing Systems 12*. MIT Press, pages 332–388, 2000.
18. L. Welling, H. Ney, A. Eiden, and C. Forbrig. Connected Digit Recognition using Statistical Template Matching. In *1995 Europ. Conf. on Speech Communication and Technology*, volume 2, Madrid, Spain, pages 1483–1486, Sept. 1995.
19. J. Wood. Invariant Pattern Recognition: A Review. *Pattern Recognition*, 29(1):1–17, January 1996.