This publication appeared as: D. Keysers, W. Macherey, H. Ney, and J. Dahmen. Adaptation in Statistical Pattern Recognition Using Tangent Vectors. In IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume 26, Number 2, pages 269-274, February 2004.

Adaptation in Statistical Pattern Recognition Using Tangent Vectors

Daniel Keysers, Wolfgang Macherey, Hermann Ney, *Member*, *IEEE*, and Jörg Dahmen

Abstract—We integrate the tangent method into a statistical framework for classification analytically and practically. The resulting consistent framework for adaptation allows us to efficiently estimate the tangent vectors representing the variability. The framework improves classification results on two real-world pattern recognition tasks from the domains handwritten character recognition and automatic speech recognition.

Index Terms—Statistical pattern recognition, adaptation, tangent vectors, linear models.

1 Introduction

Adaptation is an important topic in classification as, in many applications, recognition accuracy can be significantly improved by explicitly modeling the variability of the data. This is especially effective in cases where the training set is small. We study the use of linear representations for the variability in a statistical framework. This is related to the use of tangent vectors, which were successfully used for the recognition of handwritten digits with distance-based classifiers [16].

The main contributions of this paper are:

- to present a consistent framework for adaptation in a statistical classifier,
- to integrate the tangent vector approach into a statistical framework,
- to derive the resulting distribution analytically, and
- to evaluate the approach thoroughly using experiments on two different tasks, showing significant improvements.

The statistical framework derived in this work allows us to use tangent vectors that are the derivatives of specified transformations as well as to determine the tangent vectors from the given training data in terms of a maximum likelihood estimation. This facilitates the use of the tangent vector method for tasks where meaningful transformations of the feature vectors are not easily obtained, e.g., the transformation effects on the feature vectors of a speech signal used in automatic speech recognition.

2 Statistical Framework

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

$$x \quad \longmapsto \quad r(x) = \underset{k}{\operatorname{argmax}}\{p(k) \cdot p(x|k)\}.$$

Here, $p(k)$ is the a priori probability of class $k \in \{1, \ldots, K\}$, $p(x|k)$ 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 decision errors if the required distributions are known [4]. As they are unknown in practical situations, it is necessary to choose models for the respective distributions and estimate their parameters using the training data. We will consider:

- single Gaussian densities:

$$\begin{aligned} p(x|k) &= \mathcal{N}(x|\mu_k, \Sigma) \\ &= \det(2\pi\Sigma)^{-\frac{1}{2}} \cdot \exp\left[-\frac{1}{2}(x-\mu_k)^T\Sigma^{-1}(x-\mu_k)\right]. \end{aligned}$$

- Gaussian mixture densities:

$$p(x|k) = \sum_{i=1}^{I_k} p(i|k) \cdot \mathcal{N}(x|\mu_{ki}, \Sigma).$$

- Gaussian kernel densities:

$$p(x|k) = \frac{1}{N_k}\sum_{n=1}^{N_k} \mathcal{N}(x|x_{kn}, \Sigma). \tag{1}$$

Here, N_k represents the number of training samples of class k and x_{kn} denotes the nth reference pattern of class k. The theoretical results are derived for the case of Gaussian densities. This does not impose any restrictions as these results can be transferred to mixture or kernel densities, which can model any density function up to arbitrary precision. We assume Σ to be identical for all classes, i.e., we use variance pooling over classes. For some tasks (especially for a larger number of dimensions), we also use pooling over dimensions, i.e., $\Sigma = \alpha\sigma^2 I$ with a factor α to determine the width of the density.

Our goal is to obtain a classifier that is invariant with respect to certain transformations of the data that are known to leave the class unchanged. This goal can be addressed in different stages of the classification process: In the preprocessing step, the feature vectors can be normalized, during feature analysis, we can extract invariant features, and we can use invariant probability density functions, which are inherently related to invariant distance measures. We will concentrate here on the use of invariant probability density functions.

Invariance can also be achieved by using virtual data. This common method for creating more data than given in the training set is typically based on the invariance requirements. For example, in the experiments with optical character recognition, we use shifts in the directions of the 8-neighborhood, thus obtaining a nine-fold increase in the number of patterns. This method can be extended to the test data as well [3].

3 Tangent Vectors

3.1 Motivation

In some application areas, transformations which leave the class membership unchanged are known a priori, e.g., small affine transformations in the case of character recognition. We want the classifier to be invariant with respect to these transformations. Let $\tilde{x}(\alpha)$ denote a transformation of x depending on a parameter L-tuple $\alpha \in \mathbb{R}^L$. The set of all transformed patterns typically has highly nonlinear characteristics in pattern space. To obtain a tractable representation, we consider a linear approximation of the transformation using a Taylor expansion about $\alpha = 0$:

$$\tilde{\boldsymbol{x}}(\alpha) = x + \sum_{l=1}^{L} \alpha_l v_l + \sum_{l=1}^{L} \mathcal{O}(\alpha_l^2),$$

neglecting the terms of second order and higher. Here, the partial derivatives of the transformation $\tilde{x}$ with respect to the parameters α_l ($l = 1, \ldots, L$) are called the *tangent vectors* $v_l = \partial\tilde{\boldsymbol{x}}(\alpha)/\partial\alpha_l|_{\alpha_l=0}$, as they span the tangent subspace of the set of all transformed

• *The authors are with the Lehrstuhl für Informatik VI, Computer Science Department, RWTH Aachen-University of Technology, D-52056 Aachen, Germany.*
E-mail: {keysers, w.macherey, ney, dahmen}@informatik.rwth-aachen.de.

Manuscript received 28 Mar. 2002; revised 24 Jan. 2003; accepted 2 July 2003. Recommended for acceptance by A. Del Bimbo.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number 116174.

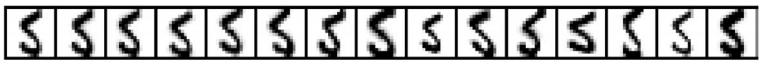


Fig. 1. Example of first-order approximation of affine transformations and line thickness. (Left to right: original image, ± horizontal translation, ± vertical translation, ± rotation, ± scale, ± diagonal deformation, ± axis deformation, and ± line thickness).

patterns at the point x. These derivatives can be efficiently calculated, e.g., using differences between slightly transformed patterns. Fig. 1 shows examples using an image of a handwritten digit and approximations of transformations. These examples illustrate the advantage of using the linear approximation as the depicted patterns (and those which result from a combination of the transformations) all lie in the same linear subspace and can, therefore, be represented by one prototype and the corresponding tangent vectors. We thus have a concise representation of the variability, where the degree of transformation is represented by a parameter vector α. This representation can be integrated into the probabilistic framework as presented in the following section.

To determine the tangent vectors $\{v_l\}$, we can use three alternatives:

(v1) compute the derivatives for the reference vector μ ($v_l = \partial\tilde{\boldsymbol{\mu}}(\alpha)/\partial\alpha_l|_{\alpha_l=0}$),

(v2) compute the derivatives for the observation vector x ($v_l = \partial\tilde{\boldsymbol{x}}(\alpha)/\partial\alpha_l|_{\alpha_l=0}$),

(v3) estimate the derivatives from the training data,

where (v1) and (v2) require prior knowledge about the transformations. How to apply (v3) will become clear with the integration of the following statistical framework which facilitates the estimation as a maximum likelihood solution.

3.2 Integration into the Probabilistic Framework: Adaptive Pattern Recognition

In adaptive pattern recognition, the distribution models are assumed to depend on an unknown adaptation parameter vector α, e.g., for rotation and scaling in image recognition [3]. The Bayesian approach to adaptation consists of integrating out the unknown parameter, which is possible in this context. We consider the case where the observations x have a Gaussian distribution with expectation μ_k and covariance matrix Σ. The extension to Gaussian mixtures or kernel densities is straightforward using maximum approximation or the expectation-maximization algorithm. The starting point is the integration

$$\begin{aligned} p(x|k) &= \int p(x,\alpha|k)d\alpha \\ &= \int p(\alpha|k)\cdot p(x|k,\alpha)d\alpha \overset{\text{model}}{=} \int p(\alpha)\cdot p(x|k,\alpha)d\alpha, \end{aligned}$$

where the distribution of the adaptation parameter set α is assumed to be independent of k. This distribution is assumed to be Gaussian with zero mean and covariance matrix equal to a multiple of the identity matrix:

$$p(\alpha) = \mathcal{N}(\alpha|0,\gamma^2 I), \tag{2}$$

where γ is a hyperparameter describing the standard deviation of the transformation parameters. The distribution of x is assumed to be Gaussian for these considerations to simplify the analytical derivation. This assumption does not imply a loss of generality as the expectation-maximization algorithm allows us to transfer the results to Gaussian mixtures or kernels, which can model arbitrarily complex distributions and are successfully used in different applications (e.g., being the standard in speech recognition).

The distribution of class k is modified for adaptation based on the first-order approximation of the transformation given by the tangent vectors $\{v_{kl}\}$:

$$\begin{aligned} p(x|k,\alpha) &= \mathcal{N}(x|\tilde{\boldsymbol{\mu}}_k(\alpha),\Sigma) \\ \tilde{\boldsymbol{\mu}}_k(\alpha) &= \mu_k + \sum_{l=1}^{L}\alpha_l v_{kl} \\ v_{kl}^T\Sigma^{-1}v_{km} &= \delta_{lm}, \end{aligned} \tag{3}$$

where $\delta_{lm} := 1$ if $l = m$ and 0 otherwise denotes the Kronecker delta. To simplify the mathematical representation, the tangent vectors are assumed to be orthonormal with respect to the global covariance matrix Σ. This does not imply a loss of generality as only the spanned subspace determines the variation modeled and it is always possible to achieve this condition using, e.g., a singular value decomposition. It is then possible to perform the integration (2) analytically by combining the exponents of the Gaussian density functions into one term of quadratic order in α using (2) and (3) and transforming this into one Gaussian density function [11]. As result we obtain the exact closed-form solution for the probability density function of the observations:

$$\begin{aligned} p(x|k) &= \mathcal{N}(x|\mu_k,\tilde{\Sigma}_k) \\ \tilde{\Sigma}_k &:= \Sigma + \gamma^2\sum_{l=1}^{L} v_{kl}v_{kl}^T \\ \tilde{\Sigma}_k^{-1} &= \Sigma^{-1} - \frac{1}{1+\frac{1}{\gamma^2}}\cdot\Sigma^{-1}\sum_{l=1}^{L} v_{kl}v_{kl}^T\Sigma^{-1}. \end{aligned} \tag{4}$$

Thus, the incorporation of tangent vectors only affects the covariance matrix, which can be interpreted as imposing a structure on Σ [11]. Note that this result does not hold for the case (v2) above—using the derivatives of the observation. In this case, the resulting distribution is not necessarily Gaussian. Fig. 2 shows an example of a resulting density that is not Gaussian. Note, furthermore, that $\det(\tilde{\Sigma}_k) = (1+\gamma^2)^L\cdot\det(\Sigma)$ (cp. [6, p. 38ff.]) which is independent of the tangent vectors and can therefore be dropped in the maximum likelihood estimation (Section 3.3).

To view the results in terms of distances, consider the exponent in $\mathcal{N}(x|\mu_k,\tilde{\Sigma}_k)$ with a covariance matrix $\Sigma = \sigma^2 I$ which is assumed to be white except for a constant factor. This is the case, for example, after application of a global whitening transform of the data. We furthermore assume $\gamma\to\infty$ here, which lets the factor $\frac{1}{1+\frac{1}{\gamma^2}}$ approach one:

$$\begin{aligned} &(x-\mu_k)^T\tilde{\Sigma}_k^{-1}(x-\mu_k) \\ &= (x-\mu_k)^T\Sigma^{-1}(x-\mu_k) - \sum_{l=1}^{L}[(x-\mu_k)^T\Sigma^{-1}v_{kl}]^2 \\ &= \frac{1}{\sigma^2}\left[(x-\mu_k)^T(x-\mu_k) - \sum_{l=1}^{L}[(x-\mu_k)^T v_{kl}]^2\right]. \end{aligned} \tag{5}$$

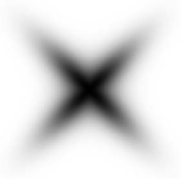


Fig. 2. The resulting density for case (v2) in a 2D example with $\Sigma = I$, $\gamma = 2$, $L = 1$, and $v_1 = \frac{1}{\|x\|}\left(\begin{smallmatrix}0&1\\1&0\end{smallmatrix}\right)x$.

The resulting exponent (5) turns out to be a modified Euclidean distance. It shows that variations along the directions of the tangent vectors are not (or less) important for classification. Note that the exponent leads to the conventional Mahalanobis distance for $\gamma \to 0$ and to the tangent distance for $\gamma \to \infty$.

3.3 Estimation of Tangent Vectors

To relax the constraint that the transformations must be known a priori, the tangent vectors can be estimated from the training data. This estimation can be formulated as a maximum likelihood approach within the presented framework. Let the training data be given by $x_{kn}, n = 1, \ldots, N_k$ training patterns of $k = 1, \ldots, K$ classes. We consider the single Gaussian model (4) with known class means μ_k and global covariance matrix Σ. For the derivation, we assume that the number L of tangent vectors is known.

We consider the log-likelihood as a function of the unknown tangent vectors $\{v_{kl}\}$:

$$\begin{aligned} F(\{v_{kl}\}) &:= \sum_{k=1}^{K} \sum_{n=1}^{N_k} \log \mathcal{N}(x_{nk}|\mu_k, \tilde{\Sigma}_k) \\ &= \frac{1}{1+\frac{1}{\gamma^2}} \cdot \sum_{k=1}^{K} \sum_{l=1}^{L} v_{kl}^T \Sigma^{-1} S_k \Sigma^{-1} v_{kl} + \text{const} \end{aligned} \tag{6}$$

with the class dependent scatter matrix

$$S_k = \sum_{n=1}^{N_k} (x_{nk} - \mu_k)(x_{nk} - \mu_k)^T.$$

Taking into account the constraints of orthonormality of the tangent vectors with respect to Σ^{-1}, we obtain the following result (cp. [6, p. 400ff.]): The class specific tangent vectors $\{v_{kl}\}$ maximizing (6) have to be chosen such that the vectors $\{\Sigma^{-1/2} v_{kl}\}$ are the eigenvectors with the largest corresponding eigenvalues of the matrix $\Sigma^{-1/2} S_k (\Sigma^{-1/2})^T$ (the dominant eigenvectors or principal components).

Using this model is equivalent to performing a global whitening transformation of the feature space (i.e., right-multiplication by $\Sigma^{-1/2}$ of all data) and then using the L principal components as tangent vectors for each class. This reduces the effect of those directions of class specific variability that contribute the most variance.

In summary, the use of estimated tangent vectors in Gaussian models consists of the following steps for each class k:

- compute the empirical mean vector μ_k,
- compute the scatter matrix S_k,
- compute $\{\Sigma^{-1/2} v_{kl}\}$ as eigenvectors with largest eigenvalues of $\Sigma^{-1/2} S_k (\Sigma^{-1/2})^T$.

4 Tasks and Experimental Results

We present experimental results using the statistical approach in combination with tangent vectors for two different real-world classification tasks, described in more detail in the following, along with the results obtained in the experiments. The tasks are from two different domains, namely, image object recognition and automatic speech recognition.

The performance of a classifier is measured by the obtained error rate (ER), i.e., the ratio of errors to the number of tests. For speech recognition, a suitable measure is the word error rate (WER). Here, the difference to the correct sentence is measured using the edit distance, defined as the minimal number of insertions (ins), deletions (del), or substitutions of words necessary to transform the correct sentence into the recognized sentence.

4.1 Image Object Recognition—USPS Corpus

Results for the domain of image object recognition were obtained on the well-known US Postal Service recognition task (USPS). It contains normalized gray-scale images of handwritten digits, taken from US postal envelopes. The images are segmented and normalized to size 16×16 pixels, yielding 256-dimensional feature vectors for the appearance based approach chosen here, where each pixel value is considered a feature. The corpus consists of 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 1a. The test set is considered to be hard (with a human error rate estimated to be 2.5 percent) and the comparably small training set makes the use of invariance methods especially helpful.

Table 1b shows a summary of results on the USPS database using the Gaussian models. The non-Gaussian data is modeled well by the use of mixture and kernel density models. Because of the good performance of the Gaussian kernel density model (1), all following experiments on USPS were based on this model, using $\Sigma = \alpha\sigma^2 I$.

In the experiments with Gaussian kernels and estimated, covariance-based tangent vectors, we computed the local scatter matrix S_{kn} using the nearest neighbors of the same class for each training vector. The experiments showed that using about 30 neighbors provides a sufficient estimate of the local covariance structure.

Fig. 3a shows the error rate with respect to γ for derivative tangents of the references and the covariance-based estimation of tangents using $L = 7$ each. It can be seen that, on this data, no significant improvement can be obtained by restricting the value of the hyperparameter γ, which controls the possible values of the transformation vectors α. This effect is most likely due to the high dimensionality of the feature space in combination with a fixed range for meaningful feature values ("black" to "white"). The strong nonlinearity of the manifolds then makes undesired solutions with high values of the parameters α very unlikely. The following experiments were therefore performed using $\gamma \to \infty$.

Another interesting factor with effect on the error rate is the number of tangent vectors used in the covariance-based approach. This dependency is depicted in Fig. 3b. It can be observed that the first four tangent vectors lead to the largest reduction in error rate, while a minimum was reached for 20 tangent vectors per kernel density. The strong decrease in error rate shows that the presented method can be effectively used to learn the class specific variability on this data set.

The effect of the three estimation methods (v1) to (v3) is indicated in Table 1a. The results show that, on this data, the covariance-based estimation of the tangent vectors (v3) leads to the same error rate as the use of the derivatives for μ (v1) using more parameters (20 instead of seven tangent vectors) or a slightly higher error rate using the same number of parameters. This result seems quite remarkable as the estimation from the data alone was able to improve results as much as the use of additional domain knowledge about the data (invariance with respect to small affine transformations and line thickness). The use of derivatives of x (v2) and the combination of (v1) and (v2) leads to further improvements.

Table 1a also contains the results obtained using additional virtual data. The use of virtual test and training data (by shifting the images 1 pixel into eight directions, keeping training and test set separated) increased the performance of the classifier further to 2.4 percent. The best result obtained using the presented approaches was with a combination of different classifiers (with varying parameters), where different test results were combined using the sum rule. This reduced the error rate further to 2.2 percent, although this last result must be considered as an effect of "training on the testing data," as the best ensemble was chosen on the basis of the test results.

Interestingly, when using a single Gaussian density, i.e., one reference per class, the error rate on the USPS corpus could be reduced from 18.6 to 5.5 percent using $L = 12$ covariance-based tangent vectors. Using only $L = 7$ tangent vectors, the result of 6.4 percent outperforms the use of the derivative, here with 11.8 percent error rate. Here, the means of the single densities are very blurred, which is a disadvantage for the derivative tangent vectors.

TABLE 1
Results for the USPS Corpus (Error Rates (ER) [%]): (a) Reported Results and (b) Results for Gaussian Models

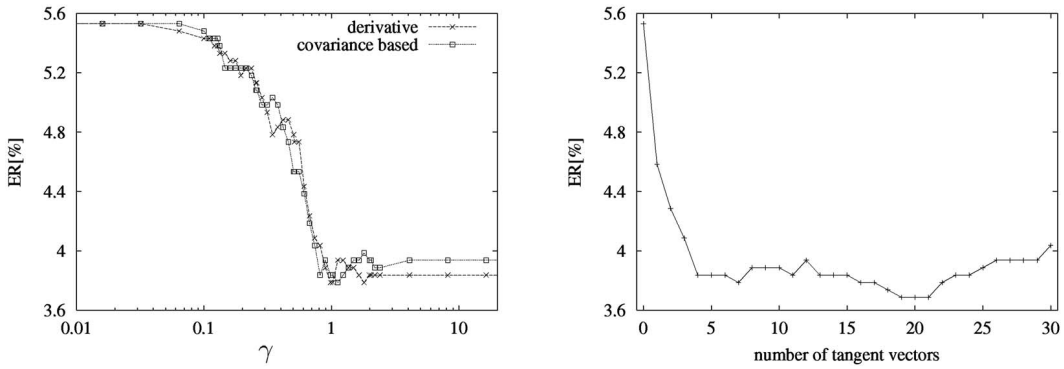
method		ER[%]
human performance [16]	[SIMARD et al. 1993]	2.5
relevance vectors [18]	[TIPPING et al. 2000]	5.1
neural net (LeNet1) [17]	[LECUN et al. 1990]	4.2
invariant support vect. [15]	[SCHÖLKOPF et al. 1998]	3.0
neural net + boosting [17]	[DRUCKER et al. 1993]	*2.6
tangent distance [16]	[SIMARD et al. 1993]	*2.5
nearest neighbor [12]		5.6
mixture densities [3]	baseline	7.2
	+ LDA + virtual data	3.4
kernel densities [12]	baseline	5.5
	+ tangent vectors (v3), $L = 7$	3.8
	(v3), $L = 20$	3.7
	(v1), $L = 7$	3.7
	(v2), $L = 7$	3.3
	(v1)+(v2), $L = 14$	3.0
	+ virtual test data	2.6
	+ virtual training data	2.4
	+ classifier combination	2.2

(a)

classifier	total # of dens.	without LDA $x \in \mathbb{R}^{16\times16}$		with LDA $x \in \mathbb{R}^{39}$
		$\Sigma = \sigma^2 I$	diag(Σ)	diag(Σ)
single Gaussian	10	18.6	19.5	12.8
Gaussian mixtures	~1,000		8.0	6.7
+ virtual data	~10,000		6.0	3.4
nearest neighbor	~7,300	5.6	6.8	7.0
+ virtual data	~65,700	4.3	5.3	3.6
Gaussian kernels	~7,300	5.5	6.3	6.5
+ virtual data	~65,700	4.2	5.1	3.4

(b)

*: training set extended with 2,418 machine-printed digits.



(a) (b)

Fig. 3. Error rate (ER) (a) as a function of tangent vector parameter standard deviation γ for $L = 7$ derivative and covariance-based tangent vectors K and (b) as a function of the number of covariance-based tangent vectors (both for USPS, kernel densities).

Using all 7,291 training patterns in a kernel density-based classifier, the result obtained without tangent model was the same as for a single density model with 12 estimated tangents (5.5 percent). In this case, the single densities with estimated tangent subspace obtain the same result as the kernel density approach using about 50 times fewer parameters.

4.2 Automatic Speech Recognition—SieTill Corpus

Experiments for automatic speech recognition were performed on the SieTill corpus [5] for telephone line recorded continuously spoken German digit strings. The corpus consists of approximately 43k spoken digits in 13k sentences for both training and test set.

The recognition system is based on whole-word Hidden Markov Models using continuous emission densities. The baseline system is characterized as follows [21]:

- vocabulary of 11 German digits, including the pronunciation variant "*zwo*,"
- gender-dependent whole-word Hidden Markov Models,
- for each gender, 214 distinct states plus one for silence,

TABLE 2
Word Error Rates (WER) on the SieTill Corpus Obtained with Tangent Distance

LDA	dns/mix	tv/mix	error rate [%]	
			del - ins	WER
no	1	0	1.17-0.83	4.59
		1	1.17-0.52	3.76
		4	0.69-1.07	3.60
	16	0	0.59-0.83	2.67
		1	0.54-0.58	2.49
		4	0.46-0.80	2.60
	128	0	0.52-0.54	2.24
		1	0.50-0.48	2.12
		4	0.55-0.49	2.13
yes	1	0	0.71-0.63	3.78
		1	0.97-0.49	3.26
		5	**0.48-0.88**	**2.70**
	16	0	0.44-0.68	2.28
		1	0.58-0.40	1.97
		4	0.38-0.55	1.97
	128	0	0.45-0.39	1.85
		1	**0.42-0.34**	**1.67**
		4	0.39-0.41	1.76

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.

- Gaussian mixture emission distributions with globally pooled diagonal covariance Σ, and
- 12 cepstral features, first derivatives, and second derivative of the first feature component.

The baseline recognizer applies maximum likelihood training using the Viterbi approximation in combination with an optional Linear Discriminant Analysis (LDA). The word error rates obtained with the baseline system for the combined recognition of both genders are summarized in Table 2 (in the lines with 0 tangent vectors (tv) per mixture (mix)). All densities of the mixtures for the states of the Hidden Markov Models were regarded as separate classes for the application of the covariance-based tangent vector estimation. The scatter matrices S_k, which are only necessary in the training phase, were trained as state specific full covariance matrices.

For single densities, the incorporation of tangent distance improved the word error rate by 18 percent relative for one tangent vector and 22 percent relative using four tangent vectors per state. In combination with LDA transformed features, the relative improvement was 14 percent for the incorporation of one tangent vector and increased to 29 percent for five tangent vectors per state. Fig. 4a depicts the word error rates on the SieTill test corpus as a function of the number of tangent vectors using single densities that were trained on LDA transformed features. For this setting, the optimal choice 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 7 percent for untransformed features and 14 percent for LDA transformed features (both at 16 dns/mix, 1 tv/mix). Consequently, a larger number of densities is able to partially compensate for the restriction that is imposed by using a globally pooled covariance matrix. The best result was obtained using 128 densities per mixture in combination with LDA transformed features and one tangent vector per state. Using this setting, the word error rate decreased from 1.85 to 1.67 percent that is a relative improvement of 5 percent. The 95 percent confidence interval for this experiment resulting in a word error rate of 1.67 percent is $[0.00\%; 1.80\%]$, which shows that the improvement is significant at the 5 percent level.

Fig. 4b depicts the word error rates for conventional training in comparison with tangent distance as a function of the number of parameters, as the incorporation of tangent vectors into the Mahalanobis distance obviously increases this number. If we compare the performance of models with the same number of parameters, we still observe that the model that includes the tangent vectors performs better.

5 Discussion

The presented tangent model is related to previous work in two fields: On the one hand, tangent vectors have been used in distance-based classifiers, where the resulting distance measure is called tangent distance. On the other hand, the resulting distributions take the form of linear Gaussian models.

Tangent distance has been successfully applied in image object recognition during the last years [12], [16], [17] and also has been recently included in textbooks [1, p. 320ff.], [4, p. 188ff.], [8, p. 423ff.] as it combines intuitive understanding and effective modeling of variability, leading to reduction of classification errors.

The subject of linear subspaces for pattern classification is treated in different contexts with different names, including principal component analysis or Karhunen-Loève transform, factor analysis, sensible principal component analysis [14], local principal component analysis [10], tangent distance [16], locally linear models [9], eigenfaces [20], etc.

Recently, [14] presented a unified view of linear Gaussian models including (sensible) principal component analysis, factor analysis, and mixtures of Gaussians with the respective

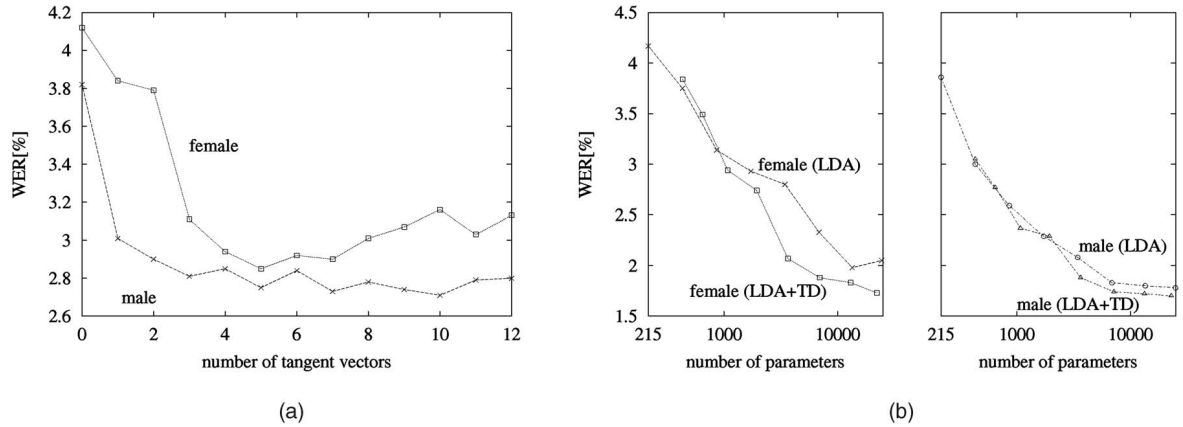


Fig. 4. (a) Word error rates (WER) as a function of the number of tangent vectors on the SieTill test corpus for single densities using ML training on LDA transformed features. (b) Comparison of word error rates for mixture densities on the SieTill test corpus using equal overall model parameter numbers.

expectation-maximization algorithms. In this work, we connect the use of tangent vectors to these models and describe a framework suitable for classification. The main addition is the treatment of the global noise covariance that is identical in the class-specific models, implying a different restriction on the covariance matrices. We consider this connection between tangent vectors and linear Gaussian models important, as the use of tangent vectors improves results on different classification tasks.

In the resulting model (4), the parameters α can be regarded as latent variables and it is therefore related to sensible principal component analysis [14] and probabilistic principal component analysis [19]. For the limiting case $\Sigma = I$, a similar result to the one presented here was derived in [7]. Note that the presented model assigns to the subspace components a weight γ which may differ from the corresponding eigenvalue, which is a main difference to subspace approximations to the full covariance matrix based on eigenvalue decomposition. In the experiments, this weight was chosen to be larger than the eigenvalues (cp. Fig. 3a). Some connections between tangent distance and linear models are already pointed out in [9], but here the authors report that they "found that the inclusion of tangent vectors did not substantially improve the performance."

The maximum likelihood estimation of the tangent vectors seems to resemble conventional principal component analysis, which minimizes the reconstruction error. But, here the projection vectors are chosen separately for each class. Furthermore, the model (4) disregards the specific variability of the patterns when determining the distance or the log-likelihood, respectively. That is, the tangent vectors span the subspace with *least* importance in the distance calculation here. In the limiting case of $\gamma \to \infty$, the effect is a class-dependent dimensionality reduction.

Note that the probabilistic interpretation of tangent distance can be used for a more reliable estimation of the parameters of a basic distribution by implicitly enriching the training set with infinitely many transformed patterns [3]. Note also that there is substantial recent work on problems related to that of determining the number of tangent vectors L automatically [2], [13], which can alternatively be achieved using cross-validation.

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 model that corresponds to the Euclidean distance or to Mahalanobis distance with diagonal covariance matrices. If full covariance matrices are used, the tangent vector approach does not increase the computational complexity.

6 Conclusion

In this paper, we presented a consistent framework for adaptation in a statistical classifier by embedding the use of tangent vectors into a probabilistic framework. The resulting model allows us to obtain transformation tolerance also if no domain knowledge about invariance properties of the feature vectors is available. The tangent vector model proved to be very effective for pattern recognition, including the combination with global feature transformations as the linear discriminant analysis.

Comparative experiments were performed on the USPS corpus for image object recognition and on the SieTill corpus for continuously spoken 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 comparable to the use of tangents based on prior knowledge. On the SieTill corpus, a relative improvement in word error rate of approximately 20 percent was achieved for single densities, and for mixture densities, we could gain a relative improvement of up to 14 percent. Incorporating the tangent vectors, we were able to significantly reduce the word error rate of our best recognition result based on maximum likelihood trained references from 1.85 to 1.67 percent.

References

[1] C.M. Bishop, *Neural Networks for Pattern Recognition.* Oxford Univ. Press 1995.

[2] C.M. Bishop, "Bayesian PCA," *Advances in Neural Information Processing Systems,* M. Kearns, S. Solla, and D. Cohn, eds., vol. 11, pp. 382-388, 1999.

[3] J. Dahmen, D. Keysers, H. Ney, and M.O. Güld, "Statistical Image Object Recognition Using Mixture Densities," *J. Math. Imaging and Vision,* vol. 14, no. 3, pp. 285-296, May 2001.

[4] R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern Classification,* second ed. New York: Wiley, 2001.

[5] T. Eisele, R. Haeb-Umbach, and D. Langmann, "A Comparative Study of Linear Feature Transformation Techniques for Automatic Speech Recognition," *Proc. Int'l Conf. Spoken Language Processing,* vol. I, pp. 252-255, Oct. 1996.

[6] K. Fukunaga, *Introduction to Statistical Pattern Recognition,* second ed. Computer Science and Scientific Computing Academic Press Inc., 1990.

[7] T. Hastie and P. Simard, "Metrics and Models for Handwritten Character Recognition," *Statistical Science,* vol. 13, no. 1, pp. 54-65, Jan. 1998.

[8] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning.* New York: Springer, 2001.

[9] G.E. Hinton, P. Dayan, and M. Revow, "Modeling the Manifolds of Images of Handwritten Digits," *IEEE Trans. Neural Networks,* vol. 8, no. 1, pp. 65-74, Jan. 1997.

[10] N. Kambhatla and T.K. Leen, "Dimension Reduction by Local Principal Component Analysis," *Neural Computation,* vol. 9, no. 7, pp. 1493-1516, 1997.

[11] D. Keysers, J. Dahmen, and H. Ney, "A Probabilistic View on Tangent Distance," *Proc. 22. DAGM Symp. Mustererkennung,* pp. 107-114, Sept. 2000.

[12] D. Keysers, W. Macherey, J. Dahmen, and H. Ney, "Learning of Variability for Invariant Statistical Pattern Recognition," *Proc. 12th European Conf. Machine Learning,* Lecture Notes in Computer Science, vol. 2167, pp. 263-275, Springer Verlag, Sept. 2001.

[13] T.P. Minka, "Automatic Choice of Dimensionality for PCA," *Advances in Neural Information Processing Systems,* T.K. Leen, T.G. Dietterich, and V. Tresp, eds., vol. 13, pp. 598-604, 2000.

[14] S. Roweis and Z. Ghahramani, "A Unifying Review of Linear Gaussian Models," *Neural Computation,* vol. 11, no. 2, pp. 305-345, 1999.

[15] B. Schölkopf, P. Simard, A. Smola, and V. Vapnik, "Prior Knowledge in Support Vector Kernels," *Advances in Neural Information Processing Systems,* M. Jordan, M. Kearns, and S. Solla, eds., vol. 10, pp. 640-646, 1998.

[16] P. Simard, Y. Le Cun, and J. Denker, "Efficient Pattern Recognition Using a New Transformation Distance," *Advances in Neural Information Processing Systems,* S. Hanson, J. Cowan, and C. Giles, eds., vol. 5, pp. 50-58, 1993.

[17] P. Simard, Y. Le Cun, J. Denker, and B. Victorri, "Transformation Invariance in Pattern Recognition—Tangent Distance and Tangent Propagation," *Proc. Neural Networks: Tricks of the Trade,* Lecture Notes in Computer Science, vol. 1524, G. Orr and K.R. Müller, eds., pp. 239-274, Springer Verlag, 1998.

[18] M.E. Tipping, "The Relevance Vector Machine," *Advances in Neural Information Processing Systems,* vol. 12, S. Solla, T. Leen, and K. Müller, eds., pp. 332-388, 2000.

[19] M.E. Tipping and C.M. Bishop, "Mixtures of Probabilistic Principal Component Analysers," *Neural Computation,* vol. 11, no. 2, pp. 443-482, 1999.

[20] M. Turk and A. Pentland, "Eigenfaces for Recognition," *J. Cognitive Neuroscience,* vol. 3, no. 1, pp. 71-86, 1991.

[21] L. Welling, H. Ney, A. Eiden, and C. Forbrig, "Connected Digit Recognition Using Statistical Template Matching," *Proc. European Conf. Speech Comm. and Technology,* vol. 2, pp. 1483-1486, Sept. 1995.