A Probabilistic View on Tangent Distance

D. Keysers, J. Dahmen, H. Ney

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

Abstract. In this paper we present a new probabilistic interpretation of tangent distance, which proved to be very effective in modeling image transformations in object recognition. Descriptions of the resulting distributions in pattern space are given for different possible models of variation, leading to a natural derivation of tangent distance. Furthermore, a possible generalization is presented and experimental results on the well known US Postal Service database are presented.

1 Introduction

Invariance of classification algorithms with respect to certain transformations plays an important role in pattern recognition. For example, in recognition of image objects like handwritten digits, invariance with respect to (small) affine variations is desired. One method which can achieve such invariance by using first order approximation of the manifolds generated by the considered transformations is known as tangent distance (TD). It was introduced by Simard et al. [14, 13] and successfully used for pattern recognition. TD and related approaches are usually seen in the context of distance based classifiers, but can as well be used in parametric classifiers [1]. For those cases, a theoretical model may be helpful, where the focus on distances can be related to the focus on distributions using the negative logarithm:

$$-\log p(x|\mu) \;=\; -\log \frac{1}{\text{norm}} \exp\left(-\frac{1}{2}d(x,\mu)\right) \;=\; \frac{1}{2}d(x,\mu) + \text{const}$$

This paper presents a novel description of the relation between a distribution respecting pattern variation and TD. In [10], a probabilistic view on subspace methods is considered, but it is only derived that the distribution of distances from the subspace has the form of a gamma distribution.

The following Section gives an overview of TD, whereas Section 3 deals with variations of the references μ respectively of the observations x and distinguishes between known derivatives of variation and cases where this information is not available. After a view on a combination of the described approaches, Section 5 gives some results and the last Section concludes the paper.

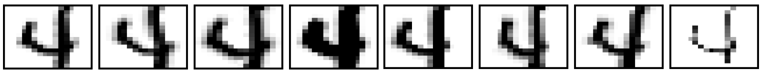

Fig. 1. Examples for tangent approximation (affine transformations and line thickness)

2 Overview of tangent distance

In 1993 Simard et al. proposed an invariant distance measure called *tangent distance*, which proved to be especially effective in the domain of digit recognition [14]. The authors observed that reasonably small transformations of certain image objects do not affect class-membership. When an image $x \in \mathbb{R}^D$ (seen as a one-dimensional vector here) is transformed (e.g. scaled and rotated) by a transformation $t(x, \alpha)$ which depends on L parameters $\alpha \in \mathbb{R}^L$ (e.g. the scaling factor and rotation angle), the set of all transformed patterns

$$M_x = \{t(x, \alpha) : \alpha \in \mathbb{R}^L\} \subset \mathbb{R}^D$$

is a manifold of at most dimension L in pattern space. The distance between two patterns can now be defined as the minimum distance between their respective manifolds, being truly invariant with respect to the L regarded transformations. As computation of this distance is a hard non-linear optimization problem and the manifolds concerned do not have an analytic expression in general, small transformations of the pattern x are approximated by a tangent subspace to the manifold M_x at the point x. This subspace is obtained by adding to x a linear combination of the vectors x_l, $l = 1, \ldots, L$ called *tangent vectors* that span the tangent subspace. The tangent vectors are the partial derivatives of $t(x, \alpha)$ with respect to α_l (therefore 'derivative' and 'direction' of variation are regarded as synonymous here). We obtain a first-order approximation of M_x, which is the subspace containing all $x_\alpha = x + \sum_l \alpha_l x_l$ for $\alpha \in \mathbb{R}^L$. The (squared) single-sided TD with tangents in x is then defined as

$$d(x, \mu) = \min_\alpha \left\{ ||x + \sum\nolimits_l \alpha_l x_l - \mu||^2 \right\}$$

The tangent vectors x_l can be computed using finite differences between the original image x and a reasonably small transformation of x [14]. Example images that were computed using tangent approximation are shown in Fig. 1 (with the original image on the left). Similarly, we can define TD using an approximation of the manifold generated by μ and a double-sided TD, where both manifolds are approximated and the distance is minimized over possible combinations of the respective parameters.

3 Probabilistic interpretation of variation

In this Section we will consider the different cases where each reference vector μ or each observation x may be subject to variations.

3.1 Known derivatives of variation in the reference

We first assume presence of a-priori knowledge about these transformations, e.g. affine transformations for images, such that the directions of variation μ_l are known. Consider a Gaussian distribution of the references with covariance matrix Σ and the first order approximation of the transformed reference

$$p(x|\mu, \alpha) = \mathcal{N}(x|\mu + \sum\nolimits_l \alpha_l \mu_l, \Sigma)$$

$$= \frac{1}{\sqrt{(2\pi)^D|\Sigma|}} \exp\left(-\frac{1}{2}\left(\mu + \sum\nolimits_l \alpha_l\mu_l - x\right)^T \Sigma^{-1}\left(\mu + \sum\nolimits_l \alpha_l\mu_l - x\right)\right)$$

Assuming independent Gaussian distribution for the α_l, $p(\alpha|\mu) = \mathcal{N}(\alpha|0,\gamma^2 I)$ (which can be justified by the central limit theorem [10]) yields

$$\begin{aligned}
p(x|\mu) &= \int p(x,\alpha|\mu)\, d\alpha = \int p(\alpha|\mu)\, p(x|\alpha,\mu)\, d\alpha \\
&\approx \max_\alpha \left\{\mathcal{N}(\alpha|0,\gamma^2 I)\, \mathcal{N}(x|\mu_\alpha,\Sigma)\right\} \qquad (1)\\
&= \max_\alpha \left\{\frac{1}{\sqrt{2\pi\gamma^2}^L}\exp\left(-\frac{1}{2\gamma^2}\sum\nolimits_l \alpha_l^2\right)\cdot\right.\\
&\left.\frac{1}{\sqrt{(2\pi)^D|\Sigma|}} \exp\left(-\frac{1}{2}\left(\mu + \sum\nolimits_l \alpha_l\mu_l - x\right)^T \Sigma^{-1}\left(\mu + \sum\nolimits_l \alpha_l\mu_l - x\right)\right)\right\} \qquad (2)
\end{aligned}$$

The use of maximum approximation in (1) is not essential. The same results (except for some constant terms) can be obtained without its application, but the calculations are somewhat more complex [8]. Expression (2) is maximized when the (double) negative logarithm is minimized, which can now be interpreted as the distance between x and μ, thus deriving an invariant distance measure (constant terms have been dropped).

$$\begin{aligned}
d(x,\mu) &:= -2\,\log p(x|\mu)\\
&\approx \min_\alpha\left\{\frac{1}{\gamma^2}\sum\nolimits_l \alpha_l^2 + \left(\mu + \sum\nolimits_l \alpha_l\mu_l - x\right)^T \Sigma^{-1}\left(\mu + \sum\nolimits_l \alpha_l\mu_l - x\right)\right\}\\
&= \min_\alpha\left\{\frac{1}{\gamma^2}\sum\nolimits_l \alpha_l^2 + (\mu - x)^T\Sigma^{-1}(\mu - x) + (\mu - x)^T\Sigma^{-1}\left(\sum\nolimits_l \alpha_l\mu_l\right)\right.\\
&\qquad \left. + \left(\sum\nolimits_l \alpha_l\mu_l\right)^T\Sigma^{-1}(\mu - x) + \left(\sum\nolimits_l \alpha_l\mu_l\right)^T\Sigma^{-1}\left(\sum\nolimits_l \alpha_l\mu_l\right)\right\}
\end{aligned}$$

Assuming orthogonality of the μ_l with respect to Σ^{-1}, that is $\mu_l^T\Sigma^{-1}\mu_{l'} = 0$ for $l \neq l'$ (which can be achieved without altering the spanned subspace using an SVD), it follows that $(\sum_l \alpha_l\mu_l)^T\Sigma^{-1}(\sum_l \alpha_l\mu_l) = \sum_l \alpha_l^2\mu_l^T\Sigma^{-1}\mu_l$. Furthermore the third and fourth term of the above sum are identical and the second term is independent of α. Therefore the expression reduces to

$$\begin{aligned}
d(x,\mu) &\approx (\mu - x)^T\Sigma^{-1}(\mu - x)\\
&\quad + \min_\alpha\left\{\sum\nolimits_l \alpha_l^2\left(\frac{1}{\gamma^2} + \mu_l^T\Sigma^{-1}\mu_l\right) + 2(\mu - x)^T\Sigma^{-1}\left(\sum\nolimits_l \alpha_l\mu_l\right)\right\}\\
&= (\mu - x)^T\Sigma^{-1}(\mu - x) - \sum\nolimits_l \frac{((\mu - x)^T\Sigma^{-1}\mu_l)^2}{\frac{1}{\gamma^2} + \mu_l^T\Sigma^{-1}\mu_l}\\
&\quad + \min_\alpha\left\{\sum\nolimits_l\left(\frac{1}{\gamma^2} + \mu_l^T\Sigma^{-1}\mu_l\right)\left(\alpha_l + \frac{(\mu - x)^T\Sigma^{-1}\mu_l}{\frac{1}{\gamma^2} + \mu_l^T\Sigma^{-1}\mu_l}\right)^2\right\} \qquad (3)\\
&= (\mu - x)^T\Sigma^{-1}(\mu - x) - \sum\nolimits_l \frac{((\mu - x)^T\Sigma^{-1}\mu_l)^2}{\frac{1}{\gamma^2} + \mu_l^T\Sigma^{-1}\mu_l}
\end{aligned}$$

where the minimization in (3) is equal to zero since it is a minimization of a sum of weighted squares. At the boundaries of the considered range for γ, $[0;\infty)$ this yields Mahalanobis distance for $\gamma \to 0$ and TD with tangents μ_l for $\gamma \to \infty$. (No gain could be obtained by restricting the value of γ.) Using the relation

$$x^T(A^{-1} + bb^T)x = x^T A^{-1} x + x^T bb^T x = x^T A^{-1} x + (b^T x)^2$$

and assuming $\gamma \to \infty$ this can be rewritten as

$$d(x,\mu) \approx (\mu - x)^T \left(\Sigma^{-1} - \sum_l \frac{(\mu_l^T \Sigma^{-1})^T (\mu_l^T \Sigma^{-1})}{\mu_l^T \Sigma^{-1} \mu_l} \right) (\mu - x) \qquad (4)$$

Eq. (4) can be regarded as assuming 'infinite' variance in the directions of the μ_l, as the inverse of the central matrix can be interpreted as covariance matrix:

$$\left(\Sigma^{-1} - \lambda \sum_l \frac{(\mu_l^T \Sigma^{-1})^T (\mu_l^T \Sigma^{-1})}{\mu_l^T \Sigma^{-1} \mu_l} \right) \left(\Sigma + \kappa \sum_l \frac{\mu_l \mu_l^T}{\mu_l^T \Sigma^{-1} \mu_l} \right)$$
$$= I - (\lambda - \kappa + \lambda\kappa) \sum_l \frac{\Sigma^{-1} \mu_l \mu_l^T}{\mu_l^T \Sigma^{-1} \mu_l}$$

The latter becomes the identity matrix I if $\lambda - \kappa + \lambda\kappa = 0$ or $\kappa = \frac{\lambda}{1-\lambda}$. Thus, as λ approaches 1 as in TD (4), κ goes to infinity, so that we can write (being aware of the fact that the inverse does not exist in $\mathrm{I\!R}^{D \times D}$):

$$p(x|\mu) = \mathcal{N}(x|\mu, \Sigma') \quad \text{with} \quad \Sigma' = \lim_{\kappa \to \infty} \left(\Sigma + \kappa \sum_l \frac{\mu_l \mu_l^T}{\mu_l^T \Sigma^{-1} \mu_l} \right)$$

The resulting distribution can be considered as a degenerate case of the normal distribution or as a normal distribution in the reduced vector space that results from the projection along the directions of the μ_l. Such a model is generally called a *linear model*, which brings about some normalization problems for the case where $\gamma \to \infty$. Hinton et al. state that such a model "is not properly normalizable", yet very useful, and refer to factor analysis as a resort [6]. This problem can be circumvented by regarding the distribution in the space originating from projection along the subspace. Note that the presented considerations can be interpreted as imposing a certain structure on the covariance matrix due to tangent distance [2].

3.2 Estimating derivatives of variation in the reference

In some cases there is no a-priori information available about the *directions* of variation of the data to be modeled, but it is known that there exists class specific variability in the data. In this case one needs to estimate the derivatives of variation for each class to be able to use the methods described above.

Given data $x_1, \ldots, x_N$, a reference μ and a covariance matrix Σ, we can apply a maximum likelihood approach to estimate the directions μ_l, assuming knowledge of the number of dimensions L to be sought for. One can show that maximizing the likelihood $\prod_n p(x_n|\mu)$ is equivalent to the maximization of the following expression with respect to the μ_l:

$$\sum_l \frac{\mu_l^T \Sigma^{-1} S \Sigma^{-1} \mu_l}{\frac{1}{\gamma^2} + \mu_l^T \Sigma^{-1} \mu_l} \overset{!}{=} \max_{\mu_1, \ldots, \mu_l}$$

with $S = \sum_n (\mu - x_n)(\mu - x_n)^T$. This is maximized when the vectors $(\Sigma^{-\frac{1}{2}})^T \mu_l$ correspond to the L eigenvectors with the largest eigenvalues of the matrix $(\Sigma^{-\frac{1}{2}})^T S \Sigma^{-\frac{1}{2}}$, its principal components. For example, assuming $\Sigma = I$ this implies using the directions of largest intra-class variance of the data. In a more general case we might consider using the global covariance matrix for Σ and the class specific covariance matrix for S, which is equivalent to performing a global whitening transformation as transformation of parameter space and then employing the L principal components of the class specific empirical covariance matrix as tangent vectors. This leads to an algorithm similar to that presented in [4], respectively within a mixture density based classifier it leads to local PCA learning [11]. Note that due to the distinction between global and class specific covariance matrix the approach we present here is inherently discriminative.

In nearest neighbor or kernel density classifiers we may be interested in a local estimation of the derivatives of variation, that is for each element x_n of the training set. Then, one approach is to use the first L principal components of the matrix $\sum_{x' \in U(x_n)} \beta(||x' - x_n||) \cdot (x' - x_n)(x' - x_n)^T$ where $U(x_n)$ is the set containing the vectors closest to x_n of the same class and $\beta(\cdot)$ is a weighting function depending on the distance of the two vectors. If $\beta(\cdot)$ is constant this yields the local subspace classifier [10]. Note that this method may not be useful for the estimation of variation in the observation during the recognition process, because then the directions need to be calculated once for every class that is hypothesized and furthermore in nearest neighbor based classifiers it leads to zero distance for all classes, if used in the straightforward manner. Therefore the following considerations deal with known variations in the observations.

3.3 Known derivatives of variation in the observation during recognition

Similar to the case of transformed references we can now consider for a given x all variations $x_\alpha = x + \sum_l \alpha_l x_l$. Since the only difference in the calculations is the replacement of the term '$+ \sum_l \alpha_l \mu_l$' by '$- \sum_l \alpha_l x_l$' in Section 3.1, we can perform exactly the same calculations, substituting μ_l with $-x_l$ and obtain (as the negation cancels out in all places)

$$d(x, \mu) = (\mu - x)^T \left(\Sigma^{-1} - \sum_l \frac{(x_l^T \Sigma^{-1})^T (x_l^T \Sigma^{-1})}{\frac{1}{\gamma^2} + x_l^T \Sigma^{-1} x_l} \right) (\mu - x) \qquad (5)$$

Note that the resulting form of the distribution cannot be expressed as a (degenerate) Gaussian here, as the matrix depends on the value of x.

3.4 Known derivatives of variation in the observation during training

One can also look at the a-priori knowledge about the data from another point of view, namely during parameter estimation, e.g. when training a Gaussian (mixture) density for recognition. In that case we might be interested in using the additional knowledge only during training for a more reliable estimation of parameters. Consider a Gaussian distribution $\mathcal{N}(x|\mu,\Sigma)$ with parameters μ and Σ to be estimated and training data $x_1,\ldots,x_N \in \mathrm{I\!R}^D$. Furthermore, we assume that the tangents $x_{n1},\ldots,x_{nL} \in \mathrm{I\!R}^D$ are given. We can now modify the maximum likelihood estimates for the parameters by distributing the weight one of each training vector x_n over "infinitely many" variations $x_{n\alpha}$ with weight $p(\alpha) = \mathcal{N}(\alpha|0,\Sigma_\alpha)$.

One can show that this has no effect on the new mean [8], i.e. $\mu_T = \mu$. Yet, the new covariance matrix does change and assuming independence and equal variance σ_α^2 for the components of α one obtains

$$\Sigma_T = \int \frac{1}{N} p(\alpha) \sum\nolimits_n (x_{n\alpha}-\mu)(x_{n\alpha}-\mu)^T \; d\alpha = \Sigma + \sigma_\alpha^2 \sum\nolimits_l \frac{1}{N} \sum\nolimits_n x_{nl} x_{nl}^T \quad (6)$$

If the resulting probabilistic models are interpreted as generative models for images, the obtained results are similar to those of Hinton et al. [7], who infer them from a variant of the neural net inspired tangent prop algorithm [13]. A similar result has also been described in [3] and for support vector machines in [12], and it is presented in a wider framework here. The estimation of parameters changes in a fundamental way, if it is assumed that TD will also be used during recognition. This has consequences for the references as well as the covariance matrix [4].

4 Combination

It is possible to combine the different approaches mentioned, e.g. combining (4) and (5) yields double-sided TD. This may be combined with (6) giving

$$d(x,\mu) = (\mu - x)^T \left(\Sigma_T^{-1} - \sum_{l=1}^{2L} \frac{(u_l^T \Sigma_T^{-1})^T (u_l^T \Sigma_T^{-1})}{u_l^T \Sigma_T^{-1} u_l} \right) (\mu - x)$$

With $\{u_1,\ldots u_{2L}\}$ being a set of vectors spanning the same subspace as the set $\{x_1,\ldots x_L,\mu_1,\ldots \mu_L\}$ with the condition $u_l^T \Sigma_T^{-1} u_{l'} = 0$ for $l \neq l'$. Since the x_l and the μ_l play essentially the same role here, and this is in turn the same as for the differences $x' - x_n$ from the Section 3.2, we might construct an even more general case, in which the first principal components of the matrix

$$\begin{aligned} &\sum\nolimits_{x' \in U(x_n)} \beta_1(||x' - x_n||)(x' - x_n)(x' - x_n)^T \\ &\qquad + \sum\nolimits_l \beta_2 x_{nl} x_{nl}^T + \beta_3 \mu_l \mu_l^T + \beta_4 \sum\nolimits_{n'} x_{n'l} x_{n'l}^T \end{aligned}$$

are used as tangent vectors for the calculation of the distance $d(x_n,\mu)$. Different settings of the coefficients $\beta_1(\cdot),\beta_2,\beta_3,\beta_4$ allow to reproduce each special case considered before, thus arriving at a valid generalization.

5 Results

All results presented here were obtained on the well known US Postal Service handwritten digits recognition task (USPS). It contains normalized greyscale images of size 16×16, divided into a training set of 7291 patterns and a test

Table 1. Summary of results for USPS
*: obtained with a training set extended by 2,400 machine-printed digits

Method	ER [%]
Human Performance [14, 13]	2.5
1-NN Classifier	5.6
This work: TD, 1-NN	3.3
TD, KD, virtual data	2.2

Method	ER [%]
Neural Net (LeNet1/4) [13]	4.2
Support Vectors [12]	3.0
Boosting [13]	*2.6
Tangent Distance [13]	*2.5

set of 2007 patterns. Reported results for this database are summarized in Table 1. Best results reported so far were obtained with an extended training set augmented with about 2,400 machine printed digits, using a nearest neighbor classifier implementing TD and a boosted neural network. In our experiments we were not able to obtain better results than 3.3% error rate with the original training set employing a 1-NN classifier with TD (affine transformations and line thickness). Using a bagged kernel density based classifier and virtual training and testing data (by shifting the images 1 pixel into 8 directions, keeping training and test set nevertheless separated), where different test results were combined using the sum rule, we were able to reduce the error rate further to 2.2%, showing the effectivity of the TD approach [9].

We also experimented with classifiers using only a single reference per class. Here, the estimation of tangent vectors in μ yielded an error rate of 6.4% for $L = 7$ (which compares favorably to 11.8% for the tangents calculated using a-priori knowledge and 18.6% for a NN without tangents) and 5.5% for $L = 12$.

To obtain results for patterns for which the derivatives of variation within each class are not known a-priori, we also carried out experiments with a reduced feature space. The patterns were transformed performing an LDA using 40 clusters of the data, yielding 39 features [1]. These features reduce the error rate without tangents from 18.6% to 12.5%. Using the estimated directions of variation this result can be improved to 8.6%. The computational complexity of the algorithms was not the key issue in the experiments but it does not impose problems, as the classification of a single observation using TD requires about one second of CPU time using all 7291 USPS training samples as references.

6 Conclusions

In this paper we presented a new probabilistic interpretation of tangent distance, deriving it from the assumption of intra-class variance. We examined different possible settings and inferred the corresponding distance measures as well as a combined representation. Tangent distance can be regarded as a structuring method for covariance matrices, assuming infinite variance in the directions of variation. Estimating the derivatives of variation amounts to local PCA if the global covariance matrix is white. The derived distance measures may be helpful in the design of classification algorithms when the considered type of variation is present in the data. The experiments carried out support our theoretical results.

Due to space limitations, some calculations were abbreviated respectively omitted. An in-depth discussion can be found in [8]. The considerations in this paper are mostly based on maximum likelihood estimation. Future work includes further investigation of the possibilities of discriminative training, taking into account the information of competing classes. One such approach that may be combined with local tangent information was presented in [5].

References

1. J. Dahmen, D. Keysers, M. O. Güld, and H. Ney. Invariant Image Object Recognition using Mixture Densities. In *Proceedings 15th International Conference on Pattern Recognition*, Barcelona, Spain, September 2000. In press.
2. J. Dahmen, D. Keysers, M. Pitz, and H. Ney. Structured Covariance Matrices for Statistical Image Object Recognition. In *22. DAGM Symposium Mustererkennung 2000*, Springer, Kiel, Germany, September 2000. This volume.
3. T. Hastie and P. Simard. Metrics and Models for Handwritten Character Recognition. *Statistical Science*, 13(1):54–65, January 1998.
4. T. Hastie, P. Simard, and E. Säckinger. Learning Prototype Models for Tangent Distance. In G. Tesauro, D. Touretzky, and T. Leen, editors, *Advances in Neural Inf. Proc. Systems*, volume 7. MIT Press, pages 999–1006, 1995.
5. T. Hastie and R. Tibshirani. Discriminative Adaptive Nearest Neighbor Classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(6):607–616, June 1996.
6. G. E. Hinton, P. Dayan, and M. Revow. Modeling the Manifolds of Images of Handwritten Digits. *IEEE Trans. on Neural Networks*, 8(1):65–74, January 1997.
7. G. E. Hinton, M. Revow, and P. Dayan. Recognizing Handwritten Digits Using Mixtures of Linear Models. In G. Tesauro, D. Touretzky, and T. Leen, editors, *Adv. in Neural Inf. Proc. Systems*, volume 7. MIT Press, pages 1015–1022, 1995.
8. D. Keysers. Approaches to Invariant Image Object Recognition. Diploma thesis, Lehrstuhl für Informatik VI, RWTH Aachen, Aachen, June 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*, Barcelona, Spain, September 2000. In press.
10. J. Laaksonen. Subspace Classifiers in Recognition of Handwritten Digits. *Acta Polytechnica Scandinavica, Mathematics, Computing and Management in Engineering Series, No. 84*, 1997. Dr. Tech. Thesis, Helsinki University of Technology.
11. 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*, London, UK, pages 497–502, 1999.
12. 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.
13. 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.
14. 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.

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