Experiments with an Extended Tangent Distance

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

Abstract

Invariance is an important aspect in image object recognition. We present results obtained with an extended tangent distance incorporated in a kernel density based Bayesian classifier to compensate for affine image variations. An image distortion model for local variations is introduced and its relationship to tangent distance is considered. The proposed classification algorithms are evaluated on databases of different domains. An excellent result of 2.2% error rate on the original USPS handwritten digits recognition task is obtained. On a database of radiographs from daily routine, best results are obtained by combining tangent distance and the proposed distortion model.

1. Introduction

Invariance with respect to certain transformations is of great interest to pattern recognition, e.g. in many cases affine transformations do not affect the class membership of image object data. SIMARD et al. [1] introduced an effective means to compensate for small affine transformations in distance based classifiers called tangent distance (TD), which led to very good results in optical character recognition (OCR). In this paper we present results of experiments with this distance in a kernel density (KD) based classifier, proposing the usage of virtual test data in addition to virtual training data. On the original United States Postal Service (USPS) OCR-database we obtain an error rate of 2.2%. Furthermore, we propose a simple but effective image distortion model (IDM) and relate it to tangent distance. The IDM considerably increased performance of the classifier with and without tangent distance on a database of medical images containing 1617 radiographs coming from daily routine.

Many approaches to invariant pattern recognition are known [2] and TD has been used in a variety of settings, including neural networks and memory based techniques like (k-) nearest neighbor algorithms (k-NN) [3], while in our experiments KD based classifiers obtained better results. A number of solutions have been proposed for efficient implementation of such algorithms, e.g. usage of hierarchical confidence refinement [4] or models for representing large subsets of the prototypes [5]. An approach motivated by deformable models, which is related to the invariance approaches covered in this paper, was proposed in [6] and tested on a face database. In comparison to TD and IDM it uses different assumptions about the allowed transformations and their cost.

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 OCR [1]. The authors observed that reasonably small transformations of certain image objects does not affect class membership. Simple distance measures like the Euclidean distance do not account for this, instead they are very sensitive to affine transformations like scaling, translation, rotation or axis deformation. When an image $x \in \mathbb{R}^{I \times J}$ 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

$$M_x = \{t(x, \alpha) : \alpha \in \mathbb{R}^L\} \subset \mathbb{R}^{I \times J} \qquad (1)$$

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. Unfortunately, computation of this distance is a hard non-linear optimization problem and the manifolds concerned generally do not have an analytic expression. Therefore, small transformations of the pattern x are approximated by a tangent subspace $\hat{M}_x$ to the manifold M_x at the point x. This subspace is obtained by adding to x a linear combination of the vectors $T_l(x), l = 1, \ldots, L$ that span the tangent subspace and are the partial derivatives of $t(x, \alpha)$ with respect to α_l. We obtain a first-order approximation of M_x

$$\hat{M}_x = \{x + \sum_{l=1}^{L} \alpha_l \cdot T_l(x) \ : \alpha \in \mathbb{R}^L\} \subset \mathbb{R}^{I \times J} \qquad (2)$$

The single-sided (SS) TD $D_{\text{TD}}(x, \mu)$ is defined as

$$D_{\text{TD}}(x, \mu) = \min_{\alpha}\{||x + \sum_{l=1}^{L} \alpha_l \cdot T_l(x) - \mu||\} \qquad (3)$$

The tangent vectors $T_l(x)$ can be computed using finite differences between the original image x and a reasonably small transformation of x [1]. Example images that were computed using (2) are shown in Fig. 1 (with the original image on the left). A double-sided (DS) TD can also be defined, where both manifolds are approximated and the distance is minimized over possible combinations of the respective parameters. It is possible to achieve a better approximation of the manifolds using an iterative procedure based on Newton's method, which is computationally more expensive.

3. The image distortion model

Although TD already compensates for small global changes, it is highly sensitive to local image transformations. We therefore propose the following image distortion model. When calculating the distance between two images x and μ local deformations are allowed, i.e. the 'best fitting' pixel in the reference image within a certain neighborhood R_{ij} is regarded instead of computing the squared error between x_{ij} and μ_{ij}. Fig. 2 shows a 1D example for the IDM (left) where individual pixel displacements are independent, in comparison to TD (right), where displacements are coupled forming an affine transformation (here scaling). The resulting distance is

$$D_{\text{IDM}}(x, \mu) = \sum_{i,j} \min_{(i',j') \in R_{ij}} \{||x_{ij} - \mu_{i'j'}|| + C_{iji'j'}\} \quad (4)$$

The cost function $C \geq 0$ represents the cost for deforming a pixel x_{ij} in the input image to a pixel $\mu_{i'j'}$ in the reference image and is introduced to compensate for the fact that in an unrestricted distortion model (i.e. with $C \equiv 0$) wanted as well as unwanted transformations can be modeled. With growing neighborhood R the admissible transformations may violate the assumption that they respect class-membership, but an appropriate choice of R leads to a significant improvement of radiograph classification even when the cost function is disregarded. To determine the cost function C, one may want to learn it from the training data or choose it empirically, e.g. by using a weighted Euclidean distance between the corresponding pixel locations. This leads to a preference of local over long-range transformations.

Figure 1. Examples for tangent approximation

4. Relating TD and IDM

It is interesting to see that the positive effects of TD and IDM are additive in some cases (see section 6). Trying to relate these two approaches it becomes clear, that one can be expressed in terms of the other. Expressing the IDM in terms of TD is difficult, when a non-zero cost function is involved (it requires additional restrictions on the values permitted for α). On the other hand generalizing the IDM leads to an expression also covering TD:

$$D_{C,\mathcal{F}}(x, \mu) = \min_{f \in \mathcal{F}}\{C(f) + \sum_{i,j} ||x_{ij} - \mu_{f(i,j)}||\} \qquad (5)$$

where $\mathcal{F} \subset (\mathrm{I\!R} \times \mathrm{I\!R})^{I \times J}$ is a class of functions assigning to each pixel its (interpolated) counterpart and $C : \mathcal{F} \to \mathrm{I\!R}^{\geq 0}$ a cost function for these assignment functions. For the IDM one has

$$\mathcal{F}_{\text{IDM}} = \{f : f(i,j) \in R_{ij}\},\ C_{\text{IDM}}(f) = \sum_{i,j} C_{ijf(i,j)} \quad (6)$$

while for TD C and $\mathcal{F}$ have the following representation:

$$\mathcal{F}_{\text{TD}} = \{f \ : \ f \text{ affine}\}, \quad C_{\text{TD}}(f) = 0 \qquad (7)$$

This general expression is an intuitive representation of a distance being invariant to arbitrary functions f of some class $\mathcal{F}$. Computing (5) may be very hard or impossible with some classes and cost functions, but TD and IDM are two examples with known solutions. (Strictly speaking (7) models the true manifold distance.) Some questions arising are e.g. which other cases are interesting in the setting of invariant pattern recognition and if one can learn the functions efficiently from training examples. For instance a model that extends the IDM naturally is to introduce a dependency between the displacements of pixels in a neighborhood, such that displacements in the same direction are cheaper than displacements in opposite directions. This leads to more complex minimization problems, which may be still efficiently solved using dynamic programming, if the number of possible displacements is small. Note that it is difficult to embed the XYI image warping approach [6] into the model (5) as the implicit XYI cost function depends on the intensity values.

5. Algorithms

We incorporated TD into a KD based classifier with Bayesian decision rule

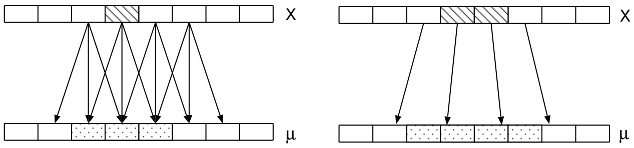

Figure 2. 1D comparison of IDM and TD

$$r(x) = \underset{c}{\operatorname{argmax}} \{p(c|x)\} = \underset{c}{\operatorname{argmax}} \{p(c)p(x|c)\} \quad (8)$$

That is we maximize the a posteriori probability for class c given an image x where the class conditional probability $p(x|c)$ is modeled by

$$p(x|c) = \frac{1}{N_c} \sum_{n=1}^{N_c} \frac{1}{A_c} \exp \left(-\frac{d(x, x_{nc})^2}{{\sigma_c}^2 \cdot \beta} \right) \quad (9)$$

with N_c being the number of training images of class c, x_{nc} the n-th reference pattern of class c, class specific standard deviation ${\sigma_c}^2$ pooled over all pixels and $d(x, x_{nc})$ being one of the proposed distance measures. To compensate for the fact that variances are usually underestimated, we multiply the estimated variances with a factor $\beta > 1$. Strictly speaking, the normalization factor A_c depends on the class c, however, the dependency is weak and therefore neglected in the experiments. Setting the a priori probability to its maximum likelihood value $p(c) = \frac{N_c}{N}$ the decision rule becomes

$$r(x) = \underset{c}{\operatorname{argmax}} \left\{ \sum_{n=1}^{N_c} \exp \left(-\frac{d(x, x_{nc})^2}{{\sigma_c}^2 \cdot \beta} \right) \right\} \quad (10)$$

Because of the exponential decay only the closest reference patterns x_{nc} have a significant contribution to the sum. The experiments show that using more than the ten closest matches does usually not change classification results. Note that this can be interpreted as a probabilistic justification for the use of k-NN based classifiers.

In order to obtain a better approximation of $p(x|c)$ the domain knowledge about invariance can be used to enrich the training set with shifted copies of the given training data. In the experiments displacements of one pixel in eight directions were used. Although the tangent distance should already compensate for shifts of that amount, this approach still leads to improvements. This is due to the fact that the two approaches model invariance differently (compare Fig. 3). As it is possible to use the knowledge about invariance for the training data by applying both tangent distance and explicit shift, this should be the case for the test data as well. Inspired by methods for combining classifiers [7] one can arrive at the following solution called virtual test sample method (VTS). When classifying a given image, shifted versions of the image are generated and independently classified. The overall result is then obtained by combining the individual results using the sum rule. An in depth discussion of this method can be found in [8].

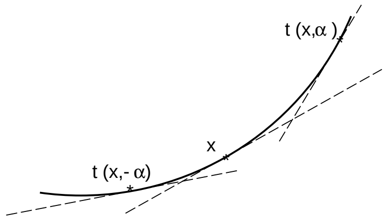

Figure 3. Tangents of shifted data in 2D

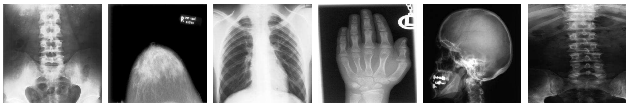

Figure 4. Example images, IRMA database

As few large differences in pixel values can mislead classifiers based on squared error distances (see e.g. [9]), it can be advisable to introduce a local threshold which limits the maximum contribution of a single pixel to the distance between two images. This is justified by a priori domain knowledge, e.g. when it is known that the patterns may be subject to artifacts that do not affect class-membership, like noise or changing scribor position in radiographs. On the other hand when looking at relatively small images of digits, one notices that e.g. changing only a few pixels can be significant for discriminating between the handwritten digits '4' and '9'. Here it can be useful to enlarge the contribution of a single pixel difference generalizing the used norm to $||x||_\gamma = (\sum_{i,j} |x_{ij}|^\gamma)^{\frac{1}{\gamma}}$ and investigating also $\gamma > 2$.

The extension of TD with an iterative Newton-type approximation was proposed in [1] and successfully used for face-recognition [9]. In our experiments we used a similar algorithm inspired by the Euler-Cauchy method used in the context of differential equations. In contrast to the Newton procedure it does not require the calculation of the actual transformation but uses the tangent approximation instead. It consists of iteratively calculating the closest point in the tangent subspace, "moving" into the corresponding direction and recalculating the tangents until convergence.

6. Experimental results

In our experiments we used three different databases. We performed a number of experiments on the widely used USPS database of handwritten digits with 7291 training and 2007 test images and validated the results on the (modified) database of the National Institute of Standards and Technology (MNIST) with 60,000 resp. 10,000 images. For comparison of performance in a different domain we also tested the algorithms on medical image data from the IRMA project (Image Retrieval in Medical Applications [10]), where the objective is to assign a given radiograph to one of the six classes abdomen, breast, chest, limbs, skull, spine (see Fig. 4). The images were scaled to a common size of 32×32 for classification without significant decrease in recognition rate. Because there were only 1617 images available, we adopted a leaving-one-out approach for cross validation, classifying each image while using the remaining 1616 as training set. After parameter adjustment the classifier was evaluated on a new set of 332 additional radiographs.

Table 1 summarizes the main results of experiments with

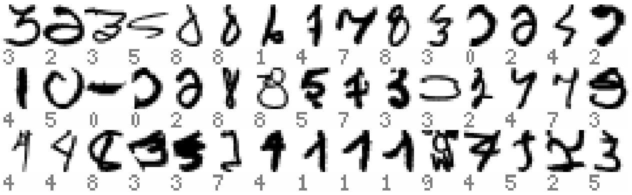

Figure 5. USPS errors (best result, 2.2% error)

the USPS database (the notation 'a-b' indicates increased number of training samples by factor a and increased number of test samples by factor b). Regardless of the chosen distance measure multiplying training and test data consistently improved classification results. The experiments showed that it is advisable to compute the tangents for the test data when computing the SSTD (1-NN performance: 3.4% vs. 3.8% for training data tangents). We chose 1-NN here as according to [3], $k = 1$ was the best choice for k-NN on USPS. Furthermore the proposed Euler-Cauchy method did not improve the overall results on this database. It produced less errors on the data misclassified by 1-NN, but more on the remaining data. This is due to the fact, that it allows larger variations in the alignment of patterns, both towards correct and incorrect reference images. Fig. 6 visualizes this effect showing the tolerance of the different distance measures with respect to a horizontal displacement. One image from each class was randomly selected and the distances to one displaced image were calculated. Multiplying the training data with tangent approximations or thinned versions of the images did not achieve lower error rates, while the usage of different norms $|| \cdot ||_\gamma$ enhanced results for the basic KD classifier, but not for the best. The variety of resulting classifiers invited the usage of classifier combination [7], and the best result could thus be improved to 2.2% error rate. Fig. 5 shows the remaining errors with their class labels. Some of the errors appear to be label mistakes, some are hard tasks even for a human, and some shapes do not appear in the training data, as is the case with the three consecutive '1's in the lower row, which were classified as '7's. Errors like the first image illustrate the limitations of distance based classifiers (there exists a training image of class five which is almost identical).

For comparison we repeated the experiment which obtained the best result on USPS on the MNIST database. Table 2 shows the results in comparison to those obtained by other groups. The asterisk indicates that the corresponding experiments were performed using a training set extended with about 2,400 machine printed digits. Note that in contrast to this approach we increase the effective size of the training set but do not actually take new data. Considering that all optimizations were done on USPS, the MNIST error rate of 1.0% is surprisingly low, which shows that the approach generalizes well and the parameters were not overfitted. Since we did not repeat all experiments for MNIST, bagging was not applied.

TD is usually applied using the seven transformations proposed in [1] (translations (2), scaling, rotation, axis-deformations (2), and line-thickness) where the first six account for affine variations. We tried a number of different tangents including projective transformations, brightness, contrast and different versions of the thickness tangent, but were not able to improve the results of the original tangents with USPS. On the other hand in a domain like medical imaging, the thickness tangent loses its a priori nature and can be replaced by a brightness tangent (here defined as a constant function over (i, j)), modeling different doses in x-ray imaging. This is reflected in the corresponding recognition rates and shows that the selection of tangents is domain dependent. Table 3 shows the results of different distance measures for the IRMA database using a region R of size 1.6×1.6 pixels, 'brightness' indicating that the tangent for line-thickness was replaced by the brightness tangent (for computational reasons we used SSTD for these experiments). Note that we were not able to obtain error rates lower than 29% using cooccurrence matrix based features [12]. The results show, that in this domain thresholding is appropriate and the improvements of TD and IDM are nearly additive. Combination of the two approaches was achieved here by replacing Euclidean distance with distortion distance (4) in the last step of distance computation, when the optimum α is already known [12]. On the other hand IDM does not improve classification for OCR (e.g. when a line distinguishing one numeral from the other is only one or two pixels thick, it is easy to erase the line completely using the IDM), which shows the domain-dependency of the distance measures involved. Using the aspect ratio of the radiographs as additional feature, the error rate dropped further from 10.3%

Table 1. Results for USPS

Distance Measure	Error Rate [%]			
	1-NN	KD	KD,9-1	KD,9-9
Euclidean	5.6	5.5	4.5	4.2
TD, SS	3.4	3.3	3.0	2.8
TD, DS	3.3	3.0	2.5	2.4

Table 2. Comparison of results for OCR

Method	Error Rate [%]	
	USPS	MNIST
Human Performance [1, 3]	2.5	0.2
Neural Net (LeNet1/LeNet4) [3]	4.2	1.1
Support Vectors [11]	3.0	0.8
Tangent Distance [3]	*2.5	1.1
Boosting [3]	*2.6	0.7
This work TD, KD, 9/9	2.4	1.0
TD, bagging	2.2	-
*obtained with extended training set		

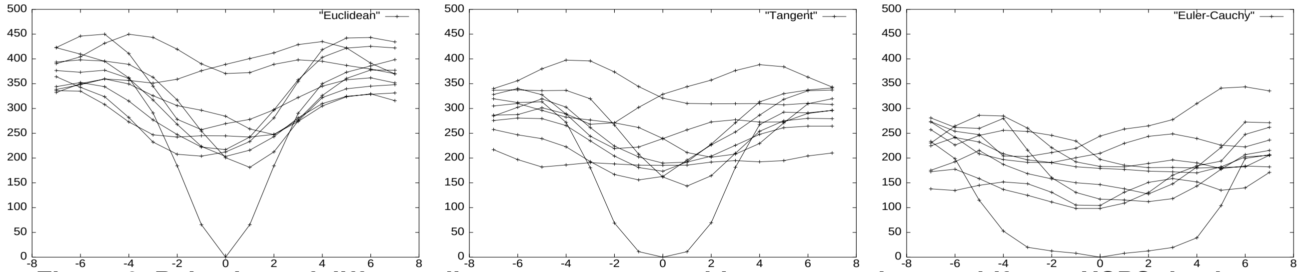

Figure 6. Behaviour of different distance measures with respect to image shifts on USPS database

Table 3. Results for IRMA database

Distance Measure	Error Rate [%]		
	1-NN	KD	KD, threshold
Euclidean	18.0	16.4	14.2
TD	15.3	14.8	13.4
IDM	16.5	14.7	13.2
TD, IDM	14.7	13.2	11.7
TD (brightness), IDM	13.5	12.9	10.3

to 8.2%. This could be improved marginally by using the Euler-Cauchy distance measure not justifying the additional amount of computation involved. Using the best parameters for the database determined by the leaving-one-out strategy, the algorithm misclassified 30 out of 332 (9.0%) new radiographs (with the training set now consisting of 1617 images) which means that an adequate generalization was achieved.

7. Conclusion

In this paper we presented a kernel density based Bayesian classifier for image object recognition. We used tangent distance to achieve tolerance with respect to affine transformations and proposed a distance measure tolerating small local variations called image distortion model. We related the two approaches and performed experiments on databases of different domains. Creating virtual training and test samples from the given data sets we obtained an excellent result of 2.2% error rate on the original USPS database. On a large database of radiographs both tangent distance and distortion model performed well and best results were obtained combining both approaches. Future work includes investigation of suitable transformations and cost functions for the generalized image distortion model (see section 4).

References

[1] P. Simard, Y. Le Cun, and J. Denker. Efficient Pattern Recognition Using a New Transformation Distance. In S. Hanson, J. Cowan, and C. Giles, editors, *Advances in Neural Inf. Proc. Systems*, volume 5, Morgan Kaufmann, San Mateo CA, pages 50–58, 1993.

[2] J. Wood. Invariant Pattern Recognition: A Review. *Pattern Recognition*, 29(1):1–17, January 1996.

[3] 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.

[4] P. Simard. Efficient Computation of Complex Distance Metrics Using Hierarchical Filtering. In J. D. Cowan, G. Tesauro, and J. Alspector, editors, *Advances in Neural Inf. Proc. Systems*, volume 6. Morgan Kaufmann Publishers, Inc., pages 168–175, 1994.

[5] 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.

[6] B. Moghaddam, C. Nastar, and A. Pentland. A Bayesian Similarity Measure for Direct Image Matching. In *Proceedings of the International Conference on Pattern Recognition, Vienna, Austria*. IEEE Computer Society Press, pages 350–358, 1996.

[7] J. Kittler, M. Hatef, and R. Duin. Combining classifiers. In *Proceedings of the International Conference on Pattern Recognition, Vienna, Austria*. IEEE Computer Society Press, pages 897–901, 1996.

[8] 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. This volume.

[9] N. Vasconcelos and A. Lippman. Multiresolution Tangent Distance for Affine-invariant Classification. In M. I. Jordan, M. J. Kearns, and S. A. Solla, editors, *Advances in Neural Inf. Proc. Systems*, volume 10. MIT Press, pages 843–849, 1998.

[10] T. Lehmann, B. Wein, J. Dahmen, J. Bredno, F. Vogelsang, and M. Kohnen. Content-based Image Retrieval in Medical Applications: A Novel Multi-step Approach. In *Procs. Int. Society for Optical Engineering (SPIE)*, volume 3972(32), pages 312–331, February 2000.

[11] 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.

[12] J. Dahmen, T. Theiner, D. Keysers, H. Ney, T. Lehmann, and B. Wein. Classification of Radiographs in the 'Image Retrieval in Medical Applications' System (IRMA). In *Proceedings of the 6th International RIAO Conference on Content-Based Multimedia Information Access, Paris, France*, pages 551–566, April 2000.