Combination of Tangent Vectors and Local Representations for Handwritten Digit Recognition

Daniel Keysers, Roberto Paredes[1], Hermann Ney, and Enrique Vidal[1]

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

[1]Instituto Tecnológico de Informática
Universidad Politécnica de Valencia
Camino de Vera s/n, 46022 Valencia, Spain
{rparedes, evidal}@iti.upv.es

Abstract. Statistical classification using tangent vectors and classification based on local features are two successful methods for various image recognition problems. These two approaches tolerate global and local transformations of the images, respectively. Tangent vectors can be used to obtain global invariance with respect to small affine transformations and line thickness, for example. On the other hand, a classifier based on local representations admits the distortion of parts of the image. From these properties, a combination of the two approaches seems very likely to improve on the results of the individual approaches. In this paper, we show the benefits of this combination by applying it to the well known USPS handwritten digits recognition task. An error rate of 2.0% is obtained, which is the best result published so far for this dataset.

1 Introduction

Transformation tolerance is a very important aspect in the classification of handwritten digits because of individual writing styles, pen properties and clutter. Among the relevant transformations we can distinguish the two cases of

- global transformations of the image, e.g. scale, rotation, slant, and
- local transformations of the image, e.g. clutter or missing parts.

These types of transformations do not change the class of the object present in the image and therefore we are interested in classifiers that can tolerate these changes, in order to improve classification accuracy. There exists a variety of ways to achieve invariance or transformation tolerance of a classifier, including e.g. normalization, extraction of invariant features and invariant distance measures.

In this work, we present two classification methods that are particularly suitable for the two types of transformations: a statistical classifier using tangent vectors for global invariance and a classifier based on the nearest neighbor

technique and local representations of the image, which tolerates local changes. Because these two methods deal with different types of transformations it seems especially useful to combine the results of the classifiers.

The combination of the classifiers is evaluated on the well known US Postal Service database (USPS), which contains segmented handwritten digits from US zip codes. There are many results for different classifiers available on this database and the combined approach presented here achieves an error rate of 2.0% on the test set, which is the best result reported so far.

2 The statistical classifier using tangent distance

First, we will describe the statistical classifier used. To classify an observation $x \in \mathbb{R}^D$, we use the Bayesian decision rule

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

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 [1]. 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 modeled using *kernel densities* in the experiments, which can be regarded as an extreme case of a mixture density model, since each training sample is interpreted as the center of a Gaussian distribution:

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

where N_k is the number of training samples of class k, x_{kn} denotes the n-th reference pattern of class k and here we assume $\Sigma = \alpha\sigma^2 I$, i.e. we use variance pooling over classes and dimensions and apply a factor α to determine the kernel width.

2.1 Overview of tangent distance

In this section, we first give an overview of an invariant distance measure, called *tangent distance* (TD), which was introduced in [2]. In the following section, we will then show how it can be effectively integrated into the statistical classifier presented above. 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.

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

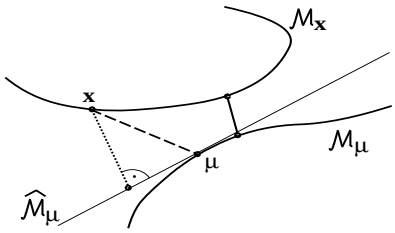

Fig. 1. Illustration of the Euclidean distance between an observation x and a reference μ (dashed line) in comparison to the distance between the corresponding manifolds (plain line). The tangent approximation of the manifold of the reference and the corresponding (one-sided) tangent distance is depicted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

affect class membership (for small α). The set of all transformed patterns now is a manifold $\mathcal{M}_x = \left\{t(x,\alpha) \ : \ \alpha \in \mathbb{R}^L\right\} \subset \mathbb{R}^D$ 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. This manifold distance is truly invariant with respect to the regarded transformations (cf. Fig. 1). However, the distance calculation between manifolds is a hard non-linear optimization problem in general. These 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 at $\alpha = 0$:

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

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

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

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 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 USPS data. These examples

Fig. 2. Example of first-order approximation of affine transformations and line thickness. (Left to right: original image, ± horizontal translation, ± vertical translation, ± rotation, ± scale, ±axis deformation, ± diagonal deformation, ± line thickness)

illustrate the advantage of TD over other distance measures, as the depicted patterns all lie in the same subspace and can therefore be represented by one prototype and the corresponding tangent vectors. 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:

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

This distance measure is also known as two-sided tangent distance (2S) [1]. To reduce the effort for determining $d_{2S}(x,\mu)$ it may be convenient to restrict the tangent subspaces to the derivatives of the reference (or the observation). The resulting distance measure is called one-sided tangent distance.

2.2 Integration into the statistical approach

The considerations presented above are based on the Euclidean distance, but equally apply when using the Mahalanobis distance in a statistical framework. The result of the integration of one-sided tangent distance into the densities is a modification of the covariance matrix of each kernel in the kernel densities [3]:

$$p(x|k) = \frac{1}{N_k} \sum_{n=1}^{N_k} \mathcal{N}(x|x_{kn}, \Sigma'_{kn}), \qquad \Sigma'_{kn} = \Sigma + \gamma^2 \sum_{l=1}^{L} \mu_{knl} \mu_{knl}^T$$

Here, the parameter γ denotes the variance of the coefficients α in the tangent subspace. The resulting distances (i.e. the values of the exponent in the Gaussian distribution) approach the conventional Mahalanobis distance for $\gamma \to 0$ and the 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_{kn}.

2.3 Virtual data

In order to obtain a better approximation of $p(x|k)$, 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, as the shift is a transformation following the true manifold, whereas the tangents are a linear approximation. As it is possible to use the knowledge about invariance for the training data by applying both tangent distance and explicit shift, this is true for the test data as well. The resulting method is called *virtual test sample method* [4]. 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.

3 The nearest neighbor classifier using local features

As the second method, the nearest neighbor (NN) paradigm is used to classify handwritten characters. To use the NN algorithm, a distance measure between two character images is needed. Usually, the solution is to represent each image as a feature vector obtained from the entire image, using the appearance-based approach as above (each pixel corresponds to one feature) or some type of feature extraction. Finally, using vector space dissimilarity measures, the distance between two character images is computed.

In the handwritten characters classification problem, there usually appear clear differences between handwritten versions of the same character. This is an important handicap to the NN classification algorithm if the feature vector is obtained from the entire image. But it is possible to find *local* parts of the characters that seem to be unchanged, that is, the distance between them is low in two handwritten version of the same character. This leads to the idea of using a local feature approach, where each character is represented by *several* feature vectors obtained from parts of the image.

In a classical classifier [1], each object for training and test is represented by a feature vector, and a discrimination rule is applied to classify a test vector. In the handwritten characters scenario, the estimation of posterior class probabilities from the whole object seems to be a difficult task, but taking local representations we obtain simpler features to learn the posterior probabilities. Moreover, we obtain a model that is invariant with respect to horizontal and vertical translations.

3.1 Extraction of local features

Many local representations have been proposed, mainly in the image database retrieval literature [5, 6]. In the present work, each image is represented by several (possibly overlapping) square windows of size $w \times w$, which correspond to a set of "local appearances" (cf. Fig. 3).

To obtain the local feature vectors from an image, a selection of windows with highly relevant and discriminative content is needed. Although a number of methods exist to detect such windows [7], most of them are not appropriate for handwritten images or they are computationally too expensive.

In this work, the grey value of the pixels is used as selection criterion. Dark pixels (with low grey value) are selected in order to determine points on the trace

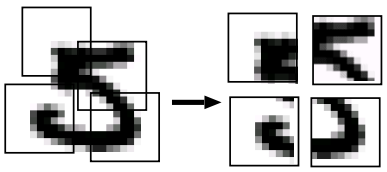

Fig. 3. Example of four local features extracted from an image of a handwritten digit.

of the handwritten character. The surrounding window of each selected pixel is used as one of the local features for the representation of the whole character.

The possibly high dimensionality of w^2 vector components is reduced using a principal component analysis on the set of all local features extracted from the training set.

3.2 Classification using local features

Given a training set, for each image of the training set a set of feature vectors is obtained. The size of these sets may be different, depending on the number of local features chosen. Each local feature vector has the same class label associated as the image it was obtained from. All these feature vectors are then joined to form a new training set.

Given a test image x, we obtain M_x feature vectors, denoted by $\{x_1, \ldots, x_{M_x}\}$. Then, to solve the problem of classification of a test object represented by local features, the sum rule is used to obtain the posterior probability of the object from the posterior probabilities of its local representations [8]:

$$r(x) = \underset{k}{\operatorname{argmax}}\, P(k|x) \approx \underset{k}{\operatorname{argmax}} \sum_{m=1}^{M_x} P(k|x_m)$$

And to model the posterior probability of each local feature, a κ-NN is used:

$$P(k|x_m) \approx \frac{v_k(x_m)}{\kappa},$$

where $v_k(x_m)$ denotes the number of votes from class k found for the feature x_m among the κ nearest neighbors of the new training set. We adopt the sum rule as an approximation for the object posterior probabilities and the k-NN estimate is used to approximate each local feature posterior probability, yielding:

$$r(x) \;=\; \underset{k}{\operatorname{argmax}} \sum_{m=1}^{M_x} \frac{v_k(x_m)}{\kappa} \;=\; \underset{k}{\operatorname{argmax}} \sum_{m=1}^{M_x} v_k(x_m) \tag{1}$$

In words, the classification procedure is summarized as follows: for each local feature of the test image, the k-nearest neighbor algorithm gives a fraction of votes to each class, which is an approximation of the posterior probability that each local feature belongs to each class. As each of the vectors obtained from the test image can be classified into a different class, a joint decision scheme is required to finally decide on a single class for the entire test image. The probabilities obtained from each local feature are combined using the sum rule to obtain the overall posterior probability for the entire image for each class. The test image is assigned to the class with highest posterior probability. According to Eq. (1), this decision corresponds to the most voted class counting all votes from all local features of the test image [8].

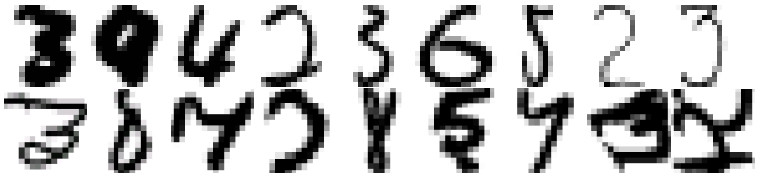

Fig. 4. Examples of digits misclassified by the local feature approach, but correctly classified by the tangent distance classifier (first row, note the variation in line thickness and affine changes) and vice versa (second row, note the missing parts and clutter).

3.3 Computational considerations

Representing objects by several local features involves a computational problem if the number of local features to represent one object is very large. The k-NN algorithm needs to compare every local feature of a test object with every local feature of every training object. This high computational cost is considerably reduced by using a fast approximate k-nearest neighbor search technique [9].

4 Experimental results on the USPS database

All the results presented here were obtained using the well known US Postal Service handwritten digits recognition corpus (USPS). It contains normalized grey scale images of size 16×16, divided into a training set of 7291 images and a test set of 2007 images. A human error rate estimated to be 2.5% shows that it is a hard recognition task. Some (difficult) examples of the test are shown in Fig. 4. Several other methods have been tried on this database and some results are included in Table 1.

Observing that the two classifiers described here led to different errors on the USPS data, this situation seemed to be especially suited for the use of classifier combination in order to improve the results [10]. For example, tangent distance is able to cope with different line thicknesses very well, while the local feature approach can tolerate missing parts (like segmentation errors) or clutter. Fig. 4 shows some of the errors, which were different between the two classifiers.

Therefore, the experimental setup was comparably simple. The best result obtained so far (2.2% error rate) was already based on classifier combination on the basis of class posterior probabilities. Hence, it was only necessary to include the results of the local feature approach (which yielded an error rate of 3.0%) in the combiner. We used the decision based on the local features with two votes, one statistical classifier with one-sided tangent distance and two statistical classifiers with two-sided tangent distance. Using majority vote as combination rule, ties were arbitrarily broken by choosing the class with the smallest class number k. With this approach, we were able to improve the result from 2.2% to 2.0%. Table 1 shows the error rates in comparison to those of other methods, which are mainly single classifier results.

Note that the improvement from 2.2% to 2.0% is not statistically significant, as there are only 2007 test samples in the test set (the 95% confidence interval

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] [2]	2.5
relevance vector machine	[TIPPING et al. 2000] [11]	5.1
neural net (LeNet1)	[LECUN et al. 1990] [12]	4.2
invariant support vectors	[SCHÖLKOPF et al. 1998] [13]	3.0
neural net + boosting	[DRUCKER et al. 1993] [12]	*2.6
tangent distance	[SIMARD et al. 1993] [2]	*2.5
nearest neighbor classifier	[14]	5.6
mixture densities	[15] baseline	7.2
	+ LDA + virtual data	3.4
(1) kernel densities	[14] tangent distance, two-sided	3.0
	+ virtual data	2.4
	+ classifier combination	**2.2**
(2) k-nearest neighbor, local representations		**3.0**
classifier combination using methods (1) and (2)		**2.0**

for the error rate on this experiment is $[1.4\%, 2.8\%]$). Furthermore, it must be admitted that these improvements seem to result from "training on the testing data". Against this impression we may state several arguments: On the one hand, only few experiments using classifier combination were performed here. Secondly, there exists no development test set for the USPS dataset. Therefore, all the results presented on this dataset (cf. e.g. Table 1) must be considered as training on the testing data to some degree and therefore a too optimistic estimation of the real error rate. This adds some fairness to the comparison. Despite these drawbacks, the presented results are interesting and important in our opinion, because the combination of two classifiers, which are able to deal with different transformations of the input (cf. Fig. 4), was able to improve on a result which was already very optimized.

5 Conclusion

In this work, the combination of two different approaches to handwritten character classification was presented. These two methods are complementary in the transformations of the images that are tolerated and thus in the sets of misclassified images. Therefore, the application of a combined classifier based on these two techniques is a suitable approach. In the experiments carried out, it was observed that the combination improves the results of the previously best classifier on the USPS corpus from 2.2% to 2.0%. Although this is not a statistically significant improvement, qualitatively, the advantages of the combination become clear when regarding Fig. 4. This shows the benefits of the applied combination, which will possibly be helpful for image classification tasks in the future.

References

1. R. Duda, P. Hart, and D. Stork. *Pattern Classification.* John Wiley & Sons, New York, 2nd edition, 2001.
2. 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 Information Processing Systems 5.* Morgan Kaufmann, pages 50–58, 1993.
3. D. Keysers, W. Macherey, J. Dahmen, and H. Ney. Learning of Variability for Invariant Statistical Pattern Recognition. In *ECML 2001, 12th European Conference on Machine Learning*, volume 2167 of *Lecture Notes in Computer Science*, Springer, Freiburg, Germany, pages 263–275, September 2001.
4. J. Dahmen, D. Keysers, and H. Ney. Combined Classification of Handwritten Digits using the 'Virtual Test Sample Method'. In *MCS 2001, 2nd International Workshop on Multiple Classifier Systems*, volume 2096 of *Lecture Notes in Computer Science*, Springer, Cambridge, UK, pages 109–118, May 2001.
5. C. Schmid and R. Mohr. Local grayvalue invariants for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(5):530–535, 1997.
6. C. Shyu, C. Brodley, A. Kak, A. Kosaka, A. Aisen, and L. Broderick. Local versus Global Features for Content-Based Image Retrieval. In *Proc. of the IEEE Workshop on Content-Based Access of Image and Video Libraries*, Santa Barbara, CA, pages 30–34, June 1998.
7. R. Deriche and G. Giraudon. A Computational Approach to Corner and Vertex Detection. *Int. Journal of Computer Vision*, 10:101–124, 1993.
8. R. Paredes, J. Perez-Cortes, A. Juan, and E. Vidal. Local Representations and a Direct Voting Scheme for Face Recognition. In *Workshop on Pattern Recognition in Information Systems*, Setúbal, Portugal, pages 71–79, July 2001.
9. S. Arya, D. Mount, N. Netanyahu, R. Silverman, and A. Wu. An optimal algorithm for approximate nearest neighbor searching fixed dimensions. *Journal of the ACM*, 45:891–923, 1998.
10. J. Kittler, M. Hatef, and R. Duin. Combining Classifiers. In *Proceedings 13th International Conference on Pattern Recognition*, Vienna, Austria, pages 897–901, August 1996.
11. M. 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.
12. P. Simard, Y. Le Cun, J. Denker, and B. Victorri. Transformation Invariance in Pattern Recognition — Tangent Distance and Tangent Propagation. In *Neural Networks: Tricks of the Trade*, volume 1524 of *Lecture Notes in Computer Science*, Springer, Heidelberg, pages 239–274, 1998.
13. B. Schölkopf, P. Simard, A. Smola, and V. Vapnik. Prior Knowledge in Support Vector Kernels. In M. Jordan, M. Kearns, and S. Solla, editors, *Advances in Neural Inf. Proc. Systems*, volume 10. MIT Press, pages 640–646, 1998.
14. 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.
15. 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.