Local Context in Non-linear Deformation Models for Handwritten Character Recognition*

Daniel Keysers, Christian Gollan, Hermann Ney
Lehrstuhl für Informatik VI – Computer Science Department
RWTH Aachen University – D-52056 Aachen, Germany
{keysers, gollan, ney}@informatik.rwth-aachen.de

Abstract

We evaluate different two-dimensional non-linear deformation models for handwritten character recognition. Starting from a true two-dimensional model, we derive pseudo-two-dimensional and zero-order deformation models. Experiments show that it is most important to include suitable representations of the local image context of each pixel to increase performance. With these methods, we achieve very competitive results across five different tasks, in particular 0.5% error rate on the MNIST task.

1 Introduction

In many object recognition tasks it is important to use suitable models of image transformation to obtain good results. The recognition of handwritten digits is a prototypical task in this context, as different writing styles lead to strongly varying appearances of the images, but at the same time the reference model for each class is well-defined. In this paper we show that two-dimensional non-linear deformation models in combination with suitable representations of the local image context of each pixel are an effective means to obtain excellent results for this task. We use five different databases and a comparison to other approaches shows that the proposed methods generalize very well.

We describe deformation models of different order: Two-dimensional (2D) hidden Markov models (HMMs) take into account the connections between the displacements of pixels in both directions of the image plane. They have been introduced in several publications, e.g. [1]. Pseudo-two-dimensional (P2D) HMMs relax the constraints in one of the dimensions, thus reducing the computational effort considerably [2]. If we further relax the constraints on the deformation grid such that the pixel displacements are independent of each other, we arrive at a zero-order model that we call image distortion model (IDM). This simple model has been introduced in the literature several times with different names, e.g. [3]. Finally, the P2DHMM can be extended to allow additional distortions.

The experiments show that the important aspect for these models is the use of local context information at the pixel level. Using this context information in the form of the image gradient and local image parts, the performance can be improved significantly, leading to state-of-the-art results.

*This work was partially funded by the DFG (Deutsche Forschungsgemeinschaft) under contract NE-572/6.

2 Classifier and deformation models

Since we investigate the influence of different deformation models on the classification accuracy, we use a simple classification setup, i.e. the well-known nearest neighbor (NN) classifier. On most databases, the 3-NN classifier was used, as other groups report good results for this choice.

The deformation models result in distance measures that are used in the NN classifier, where the test image $A = \{a_{ij}\}, i=1,\ldots,I, j=1,\ldots,J$ is explained by a suitable deformation of the reference image $B = \{b_{xy}\}, x = 1,\ldots,X, y = 1,\ldots,Y$. Here, the image pixels take U-dimensional values $a_{ij}, b_{xy} \in \mathbb{R}^U$. We now want to determine an image deformation mapping $(x_{11}^{IJ}, y_{11}^{IJ}) : (i,j) \mapsto (x_{ij}, y_{ij})$ that results in the distorted reference image $B_{(x_{11}^{IJ}, y_{11}^{IJ})} = \{b_{x_{ij}y_{ij}}\}$. The resulting image difference cost function is then defined as

$$C\left(A, B, (x_{11}^{IJ}, y_{11}^{IJ})\right) = \sum_{i,j} \sum_{u} ||a_{ij}^u, b_{x_{ij}y_{ij}}^u||^2,$$

i.e. by summing up the local pixel distances. Here, the squared Euclidean pixel distance is used, which means that the result is the squared Euclidean distance between the test image and the distorted reference image.

In addition to the image difference cost C, costs R for the deformation are introduced. The structure of the cost functions for the image deformation determines the type of the deformation model. We can distinguish absolute and relative deformation cost functions. Absolute cost functions depend only on the absolute pixel displacement:

$$R_{f^a}^a\left((x_{11}^{IJ}, y_{11}^{IJ})\right) = \sum_{i,j} f^a(x_{ij} - i, y_{ij} - j)$$

On the other hand, relative cost functions depend on the relative displacement of neighboring pixels:

$$R_{f^r}^r\left((x_{11}^{IJ}, y_{11}^{IJ})\right) = \sum_{i,j} f^r(x_{ij} - (x_{i-1j}+1), y_{ij} - y_{i-1j}, x_{ij} - x_{ij-1}, y_{ij} - (y_{ij-1}+1))$$

Here, f^a and f^r determine the model structure as mentioned above. The distance measure is determined by mini-

mizing the costs over the possible deformation mappings:

$$D(A,B) = \min_{(x_{11}^{IJ}, y_{11}^{IJ})} \left\{ C\left(A, B, (x_{11}^{IJ}, y_{11}^{IJ})\right) + \alpha R\left((x_{11}^{IJ}, y_{11}^{IJ})\right) \right\},$$

where α is the deformation cost weight and $R = R^r + R^a$ is the sum of relative and absolute deformation costs. Additionally, the border constraints $x_{1j} = 1, x_{Ij} = X, y_{i1} = 1, x_{iJ} = Y$ are applied. To soften these constraints the images can be extended by e.g. 3 pixels at the borders.

Two-dimensional HMM. The 2DHMM is an extension to two dimensions of the (0,1,2)- or (loop, jump, skip)-HMM that is frequently used e.g. in speech recognition. It results from using the relative cost function

$$f^r(\Delta_i^x, \Delta_i^y, \Delta_j^x, \Delta_j^y) = \begin{cases} \infty & \max(|\Delta_i^x|, |\Delta_i^y|, |\Delta_j^x|, |\Delta_j^y|) > 1 \\ 0 & \text{otherwise} \end{cases}$$

This cost function ensures monotonicity (no backward steps) and continuity (no large jumps) of the displacement grid. Here, instead of using the cost value 0, we can also use a function depending on the relative displacements. In addition, we may also use an absolute cost function

$$f^a(\Delta^x, \Delta^y) = \begin{cases} \infty & \max(|\Delta^x|, |\Delta^y|) > w \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

with a warp range w (e.g. $w = 2$), which restricts the absolute global deformation. The minimization of this true 2D model is NP-complete [4], and therefore approximation algorithms are used as e.g. dynamic programming with beam-search [1], or simulated annealing.

Pseudo-two-dimensional HMM. The P2DHMM [2] is obtained from the 2DHMM by neglecting the dependencies between pixels of neighboring image columns, using:

$$f^r(\Delta_i^x, \Delta_i^y, \Delta_j^x, \Delta_j^y) = \begin{cases} \infty & \max(|\Delta_i^x|, |\Delta_j^y|) > 1 \vee |\Delta_j^x| > 0 \\ 0 & \text{otherwise} \end{cases}$$

Here, the relative displacement Δ_i^y in the y direction between neighboring columns is neglected, and all pixels from one column are mapped onto the same target column. Again, we can use a function depending on the relative displacements and an absolute warp range as in Eq. (1).

This model is computationally equivalent to a one-dimensional HMM and therefore the minimization process can be solved in polynomial time. Nevertheless, the computational effort can be substantial and methods like beam search can be used to speed up the minimization.

Image Distortion Model. The IDM is a zero-order model, i.e. local dependencies in the displacement grid are neglected. Consequentially, it results from using no relative cost function (i.e. $f^a = 0$) but only using a global warp range as an absolute cost function as in Eq. (1). Figure 1 shows an example mapping of a pixel within the IDM. Since the dependencies are neglected, the minimization process for the IDM is computationally inexpensive. In comparison with the Euclidean distance it needs approximately a factor of $(2w+1)^2$ more in computation time.

Pseudo-two-dimensional HM distortion model. The P2DHMDM is a combination of the P2DHMM and the IDM. It results from the P2DHMM when we allow additional distortions from the columns that are matched by an additional pixel, where these distortions are independent of each other. Figure 2 shows an example mapping allowed under the restrictions of the P2DHMDM.

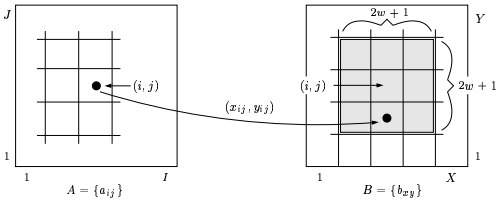


Figure 1. IDM mappings, cp. Eq. (1).

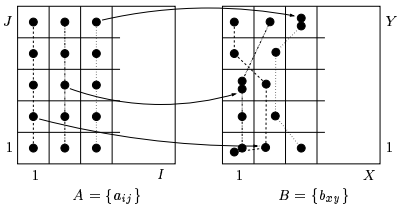


Figure 2. Example P2DHMDM mapping.

3 Using local image context

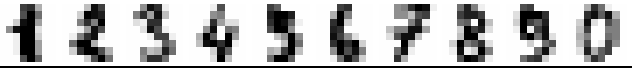
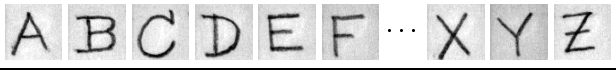
Using the above models, experiments on the USPS corpus (Section 4) showed that only a comparatively small improvement from 5.6% to 4.0% was possible when using the pixel values directly. Furthermore, the improvements depended strongly on the choice and weight of the deformation cost function for the allowed relative displacements. An analysis showed that some wanted deformations but also unwanted deformations changing the class membership of the images were modeled. This behavior can be restricted by including the local image context of each pixel in an appropriate way, with the result that pixel values are vectors.

Derivatives as context. One straight forward way to include the local image context is to use derivatives of the image values with respect to the image coordinates as computed by the horizontal and vertical Sobel filter. These values have the additional advantage of invariance with respect to the absolute image brightness.

Now the question arises of how to weight the importance of the context information with respect to the image gray values. Figure 3 shows the error rate on the USPS corpus (using the P2DHMM) with respect to the relative weight of the gradient image (a relative weight of 1 means that only the gradient information is used). From the graph it is clear that best results are obtained when using only the gradient information. The additional use of the second derivative lead to only small improvements in first experiments.

Sub images as context. A second way to include the local image context is to use local sub images that are extracted around the regarded pixel, e.g. of size 3×3 pixels. If these contexts are extracted from the gradient images, the value of an image pixel is a vector of dimension $2 \cdot 3 \cdot 3 = 18$.

Table 1. Corpus and image sizes and example images.

name	example images	size	# train	# test
USPS		16 × 16	7 291	2 007
UCI		8 × 8	3 823	1 797
MCEDAR		8 × 8	11 000	2 711
MNIST		28 × 28	60 000	10 000
ETL6A		64 × 63	15 600	13 000

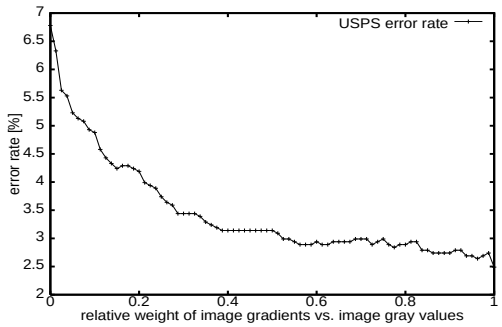


Figure 3. USPS error vs. gradient weight.

4 Databases and results

All results presented here were obtained using 3×3 sub images of the horizontal and vertical gradient images only, i.e. 18-dimensional vectors as pixel values, as these settings showed the overall best performance among those investigated. Figure 4 shows the process of the feature extraction: The horizontal and vertical gradient images are calculated using the Sobel filter, then the local 3×3 context is extracted in the gradient images and the values are stacked onto each other to form the pixel-level feature vector. All experiments were performed using a k-NN classifier with $k \in \{1, 3\}$. To speed up the classification process, a preselection of the e.g. 500 best fitting references based on the Euclidean distance was performed in some of the experiments [5]. The software used for the experiments is available for download[1].

An overview of the used corpora is shown in Table 1. For each database some example images are shown along with the sizes of training and test sets and the sizes of the images. Reference results are shown in Table 2 along with the results obtained using the methods presented in this paper. It can be observed that the results are state-of-the-art and even improve on the best published error rates for three of the five corpora. In the following paragraphs we shortly summarize special results for each of the databases used.

USPS. On the US postal service database, the P2DHMDM gave the best results. Some experiments were performed for training of prototypes with the presented models. Figure 5 shows the mean images for all ten classes and the prototypes learned using the deformation model. The error rates show that the P2DHMDM performs better using the learned prototypes and interestingly using only one prototype per class, the error rate is as low as 4.9%. The learned prototypes appear much less blurred than the means, as the variation is compensated by the non-linear deformation model. On the other hand the corresponding mean images perform better if no deformation is used.

[1] http://www-i6.informatik.rwth-aachen.de/~gollan/w2d.html

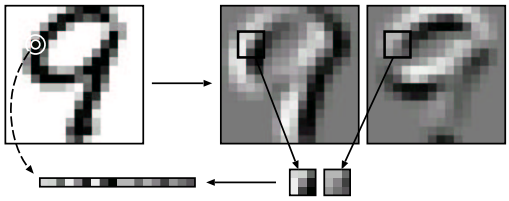
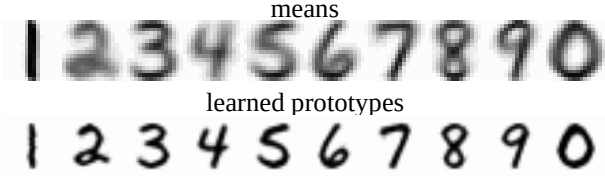


Figure 4. Extraction of local image context.

UCI. On the University of California, Irvine, optical digits corpus, a scaling to 16×16 pixels using spline interpolation was performed. Here, the IDM performed as good as the more complex P2DHMDM.

MCEDAR. On the Modified CEDAR (Center of Excellence for Document Analysis and Recognition) task, again the images were scaled to 16×16 pixels using splines. Here, the P2DHMDM performed slightly better than the IDM.

MNIST. On the Modified National Institute of Standards and Technology task, due to the good results of the IDM on the other databases and the lower complexity, only the IDM



distance	means	learned prototypes
no deformation	18.6	26.1
P2DHMDM	25.3	4.9

Figure 5. USPS prototypes / error rates [%].

Table 2. Error rates for the different corpora.

method		ER[%]
USPS		
Euclidean distance, 1-NN		5.6
2DHMM, 1-NN	this work	2.7
extended tangent distance	[3]	2.4
IDM, 1-NN	this work	2.4
extended support vectors	[6]	2.2
local features + tangent distance	[7]	2.0
P2DHMDM, 3-NN	this work	**1.9**
UCI		
Euclidean distance, 1-NN		2.0
PCA mixture	[8]	1.5
P2DHMDM / IDM, 1-NN	this work	**0.8**
MCEDAR		
factor analysis	[9]	4.7
probabilistic PCA	[10]	4.6
IDM, 3-NN	this work	3.5
P2DHMDM, 3-NN	this work	**3.3**
MNIST		
deslant, Euclidean distance, k-NN	[11]	2.4
extended tangent distance	[3]	1.0
distortions, neural net, boosting	[11]	0.7
shape context matching, 3-NN	[12]	0.6
invariant support vector machine	[13]	0.6
IDM, 3-NN	this work	**0.5**
distortions+, neural net	[14]	0.4
ETL6A		
Euclidean distance, 1-NN		4.5
piece-wise linear 2D-HMM	[15]	0.9
Eigen-deformations	[16]	0.5
IDM, 3-NN	this work	**0.5**

was tested. It resulted in 54 errors as compared to the 63 errors reported in [12] and 56 reported in [13].

ETL6A. The Electrotechnical Laboratory, National Institute of Advanced Industrial Science and Technology, Japan, 6A sub corpus contains Latin uppercase letters of 26 classes that were scaled down to 16×16 pixels. The proposed deformation models, which are similar to the methods used in [15, 16], also obtained very good results, here.

Comparison of different models. Due to space limitations, we cannot give a complete list of reference results for the different models here. A detailed discussion can be found in [5]. In the experiments, some general results could be observed. For all of the models, the performance increased significantly with the use of local image context. This increase was greater for the simpler models, leading to the conclusion that the context information can compensate for the neglected restrictions. One of the reasons for the fact that the more complex models did not outperform the simpler models is the following: Due to the computational complexity of the minimization, approximation methods had to be applied and for the more complex models usually a smaller number of images was preselected using the Euclidean distance to reduce the computational effort.

5 Conclusion

We presented different non-linear deformation models and introduced the P2DHMDM as an extension. The most important aspect when applying these methods is the inclusion of local image context information, for which we propose the gradient and small sub images. Using the context information, very good results can be achieved even with simple deformation models. The experiments gave new best results on three of the five databases used and were very competitive on the remaining two. The excellent overall results show the general applicability of the methods, which is also true for medical images [17]. Aspects to be addressed in future work include other methods of context extraction, as e.g. PCA, the use of more prototypes per class and the extension to the recognition of continuous script.

References

[1] S. Uchida, H. Sakoe. A monotonic and continuous two-dimensional warping based on dynamic programming. In *Proc. ICPR*, Brisbane, Australia, pp. 521–524, Aug. 1998.

[2] S. Kuo, O. Agazzi. Keyword Spotting in Poorly Printed Documents using Pseudo 2-D Hidden Markov Models. *IEEE Trans. PAMI*, 16(8):842–848, Aug. 1994.

[3] D. Keysers, J. Dahmen, T. Theiner, H. Ney. Experiments with an Extended Tangent Distance. In *Proc. ICPR*, Barcelona, Spain, pp. 38–42, Sept. 2000.

[4] D. Keysers, W. Unger. Elastic Image Matching is NP-complete. *Pattern Recog. Lett.*, 24(1–3):445–453, Jan. 2003.

[5] C. Gollan. Nichtlineare Verformungsmodelle für die Bilderkennung (in German). Diploma thesis, Lehrstuhl für Informatik VI, RWTH Aachen University, Germany, Sept. 2003.

[6] J.X. Dong, A. Krzyzak, C.Y. Suen. A Practical SMO Algorithm. In *Proc. ICPR*, Quebec City, Canada, Aug. 2002.

[7] D. Keysers, R. Paredes, H. Ney, E. Vidal. Combination of Tangent Vectors and Local Representations for Handwritten Digit Recognition. In *Int. Workshop Stat. Pattern Recognition*, Windsor, Ontario, Canada, pp. 538–547, Aug. 2002.

[8] H.-J. Kim, D. Kim, S.Y. Bang. A Numeral Character Recognition using the PCA Mixture Model. *Pattern Recog. Lett.*, 23(1–3):103–111, Jan. 2002.

[9] G.E. Hinton, P. Dayan, M. Revow. Modeling the Manifolds of Images of Handwritten Digits. *IEEE Trans. Neural Networks*, 8(1):65–74, Jan. 1997.

[10] M. Tipping, C. Bishop. Probabilistic Principal Component Analysis. *J. Royal Stat. Soc. (B)*, 61(3):611–622, 1999.

[11] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner. Gradient-Based Learning Applied to Document Recognition. *Proc. of the IEEE*, 86(11):2278–2324, Nov. 1998.

[12] S. Belongie, J. Malik, J. Puzicha. Shape Context: A New Descriptor for Shape Matching and Object Recognition. In *NIPS 13*. MIT Press, pp. 831–837, 2001.

[13] D. DeCoste, B. Schölkopf. Training Invariant Support Vector Machines. *Machine Learning*, 46:161–190, 2002.

[14] P. Simard. Best Practices for Convolutional Neural Networks Applied to Visual Document Analysis. In *Proc. ICDAR*, Edinburgh, Scotland, pp. 958–962, Aug. 2003.

[15] S. Uchida, H. Sakoe. Handwritten character recognition using elastic matching based on a class-dependent deformation model. In *Proc. ICDAR*, Edinburgh, Scotland, UK, pp. 163–167, Aug. 2003.

[16] S. Uchida, H. Sakoe. Eigen-Deformations for Elastic Matching based Handwritten Character Recognition. *Pattern Recognition*, 36(9):2031–2040, Sept. 2003.

[17] D. Keysers, C. Gollan, H. Ney. Classification of Medical Images using Non-linear Distortion Models. In *Bildverarbeitung für die Medizin*, Berlin, Germany, pp. 366–370, March 2004.