Deformation Models for Image Recognition

Daniel Keysers, Thomas Deselaers, *Student Member*, *IEEE*,
Christian Gollan, and Hermann Ney, *Member*, *IEEE*

Abstract—We present the application of different nonlinear image deformation models to the task of image recognition. The deformation models are especially suited for local changes as they often occur in the presence of image object variability. We show that, among the discussed models, there is one approach that combines simplicity of implementation, low-computational complexity, and highly competitive performance across various real-world image recognition tasks. We show experimentally that the model performs very well for four different handwritten digit recognition tasks and for the classification of medical images, thus showing high generalization capacity. In particular, an error rate of 0.54 percent on the MNIST benchmark is achieved, as well as the lowest reported error rate, specifically 12.6 percent, in the 2005 international ImageCLEF evaluation of medical image categorization.

Index Terms—Image matching, image alignment, character recognition, medical image categorization.

✦

---

1 INTRODUCTION

IMAGE matching is an important step in many image recognition tasks. Especially in the presence of high image variability, classification by flexible matching of an image to given references is one of the most promising approaches to achieve low error rates. In this paper, we discuss and compare several models for image matching in the context of recognition and show that they lead to excellent results across different tasks.

Numerous deformation models of different complexity have been discussed in the literature, but they are usually not directly compared. Models of second, first, and zeroth order exist. However, a large number of problems have not yet been addressed, specifically:

- What is the trade-off between algorithmic complexity and classification performance?
- What is the effect of pixel context in matching algorithms?
- How important are true 2D approaches?

Here, we provide a consistent performance comparison on real-world tasks and also compare the computational complexities. We denote by *order 2* those models for which the displacement of a pixel depends on the displacement of its neighboring pixels in both directions of the two-dimensional image grid. Analogously, we say a model is of *order 1* if this dependency is reduced to neighbors along one dimension and of *order 0* if no such dependency exists. Starting from a true two-dimensional model, we proceed to discuss pseudo-two-dimensional models (of order one) and

zero-order deformation models. In the past, it was unclear to which extent complex models with many constraints on the matchings are necessary for good recognition results. It is to be expected and will be confirmed by the experiments that—for the tasks investigated—complex models are not necessary; a simple model that incorporates a suitable representation of the local image context is sufficient.

The main objectives of this paper are:

- To show that a conceptually simple nonlinear model of image variability leads to consistently high performance in several real-world image recognition tasks and might therefore be considered as a baseline method for various recognition tasks.
- To show that the straightforward paradigm of appearance-based image classification with appropriate models of variability leads to very competitive results in the domain of handwritten character recognition.
- To directly compare different nonlinear models and to experimentally confirm that the use of less restrictive two-dimensional constraints in image matching can be compensated by using local image context at the pixel level.

In this paper, we call the process of assigning one out of K class labels to an image *recognition* or *classification*. A *deformation* of an image is the application of a two-dimensional transformation of the image plane, e.g., a small rotation or a shift of a small part of the image. The *matching* of two images consists of finding the optimal deformation from a set of allowed deformations in the sense that it results in the smallest distance between the deformed reference image and the observed test image. The *context* of a pixel refers to the values of pixels in a neighborhood of that pixel and quantities derived from these, e.g., gradient values.

2 RELATED WORK ON IMAGE MATCHING

There is a large amount of literature dealing with the application of matching to computer vision and pattern recognition tasks. In this paper, we focus on the local deformation of images and, thus, do not discuss the global

---

- *D. Keysers is with the German Research Center for Artificial Intelligence (DFKI GmbH), Image Understanding and Pattern Recognition Group, D-67663 Kaiserslautern, Germany. E-mail: daniel.keysers@dfki.de.*
- *T. Deselaers, C. Gollan, and H. Ney are with the Lehrstuhl für Informatik 6, Computer Science Department, RWTH Aachen University, D-52056 Aachen, Germany.*
  E-mail: {deselaers, gollan, ney}@informatik.rwth-aachen.de.

alignment of images, which, with respect to affine transformations, for example, is a problem that is complementary to the one discussed here. Note that matching is often done after segmentation or contour extraction, which is inherently a two-stage procedure. We propose using the appearance-based approach in which no intermediate segmentation errors can occur. We briefly discuss the most relevant related work here.

Uchida and Sakoe [52] present a dynamic programming algorithm for the true two-dimensional image deformation model which takes into account continuity and monotonicity constraints. The exponential computational complexity is reduced by finding possibly suboptimal solutions using beam search. In [53], Uchida and Sakoe investigate the use of class-specific deformations for handwritten character recognition using three different elastic matching techniques. These eigen-deformations improve the recognition results considerably, but no pixel context is taken into account. Recently, the authors presented a thorough comparison and classification of different warping algorithms in [54].

Wiskott et al. [57] represent face images by elastic graphs which have node labels representing the local texture information as computed by a set of Gabor filters. These filters represent local information comparable to the Sobel values used in this work, but a richer descriptor on a sparse grid is used.

Belongie et al. [2] describe a method for image matching that is based on representations of the local image context called "shape contexts" which are extracted at edge points. An assignment between these points is determined using the Hungarian algorithm and the image is matched using thin-plate splines, which is iterated until convergence. This method also uses a richer descriptor than the method presented here and relies on edge detection first; it is also computationally much more complex.

Levin and Pieraccini [34] extend the one-dimensional dynamic time warping algorithm to two dimensions and note that the algorithm has exponential complexity. They then propose restricting the problem by assuming independence between vertical and horizontal displacement and thus arrive at a model that is essentially a pseudo-two-dimensional (P2D) model.

Kuo and Agazzi [30] also describe the use of pseudo-two-dimensional hidden Markov models (HMMs) in the domain of document processing. They use such a model to detect the occurrences of keywords in images of documents that are subject to noise.

The use of iterative matching using coupled one-dimensional HMMs is proposed for two-dimensional matching by Perronnin et al. [42], but no pixel context is considered.

Image registration is a concept that is often applied in medical applications and is connected to the methods discussed here by the inherent optimization or matching process. A prominent example of a registration algorithm is the one by Viola and Wells [56], but the literature with respect to this subject is vast. In their review of image warping methods, Glasbey and Mardia [15] give a summary of various methods for image registration or warping and state that the fundamental trade-off is between the smoothness of the image transform and the closeness of the achieved match. The major difference between image registration and matching for recognition is that, in registration, it is known that the two images should match. The trade-off in image registration is between a good match and high distortion of the image. In matching for image recognition, on the other hand, we do not know in advance if the two images should result in a good match. Thus, we are faced with a different trade-off here: We are interested in a matching that allows us to compensate for geometric intraclass variation but at the same time retains interclass variation even under the optimal matching allowed within the model. In terms of distances, we have to balance wanted small distances for images of the same class (which can be achieved by adding more flexibility to the matching process) with the unwanted result of generally small distances even for images of different classes (which requires limiting the flexibility of the matching process).
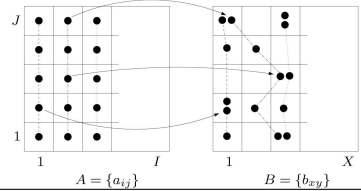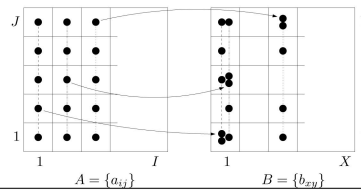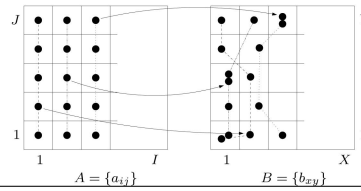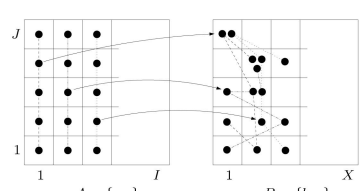
The model that we present as the "image distortion model" was presented earlier in several other works under different names: For example, it is presented by Burr in [6]; in [41], Mori et al. refer to it as "shift similarity;" Umeda presents it as "locally maximized correlation" in [55]; and, recently, it was referenced as "local perturbation" by Uchida and Sakoe in [54]. In contrast to these references, we emphasize the use of local context within this model, which greatly improves recognition results.

In the past, matching for handwritten digit recognition has also been approached as a problem of matching a one-dimensional structure in two-dimensional images (the production of handwritten symbols by drawing with a pen implies that there is an underlying one-dimensional structure), e.g., [5], [20]. However, in the current literature, it seems that, for offline recognition, two-dimensional models perform considerably better, although outlines are still used directly [7] and indirectly [2].

Summarizing, these previous methods can roughly be categorized by 1) the representation of the underlying image parts being matched, 2) the allowed deformations, and 3) the classification step.

1. *Representation of the underlying image parts.* Local texture descriptors (roughly comparable to the Sobel filters used here) are used in [57]. Belongie et al. [2] use spatial histograms over edge points to represent shapes. Some works directly use the gray values of pixels for the matching [43], [52], [54], [42]. For OCR applications, some methods use outlines [7], [2] or rely on an underlying one-dimensional structure of the patterns [5], [20].

2. *Allowed deformations.* The bunchgraph matching technique allows for rotation, scale, translation, and small local displacements [57]. In the shape context method, the allowed transformations are described by thin-plate-spline transformations. The one-dimensional matching [6], [41], [55], [54] that we call "image distortion model" allows for nonlinear local displacements. The methods described in [34], [30] allow for the same deformations as the P2DHMM model presented here, whereas the method presented in [42] iteratively improves these deformations by alternating horizontal and vertical direction.

3. *Classification step.* The methods presented in [2], [43], [52], [54] use the nearest neighbor or nearest prototype classification rule for the classification of images and the result of the matching process is used to obtain an appropriate distance function. Most of the experiments performed in this work similarly use the nearest neighbor decision rule as this allows us to

TABLE 1
Overview of the Constraints Imposed on the Image Deformation Mappings for the Different Deformation Models

2DW	2-Dimensional Warping (second-order), complete 2D constraints, minimization NP-complete $x_{1j} = 1, x_{Ij} = X, y_{i1} = 1, y_{iJ} = Y,$ $x_{i+1,j} - x_{ij} \in \{0,1,2\}, x_{i,j+1} - x_{ij} \in \{-1,0,1\},$ $y_{i,j+1} - y_{ij} \in \{0,1,2\}, y_{i+1,j} - y_{ij} \in \{-1,0,1\}$	
P2DHMM	Pseudo 2-Dimensional Hidden Markov Model (first-order), match columns onto columns, columns are independent $x_{1j} = 1, x_{Ij} = X, y_{i1} = 1, y_{iJ} = Y,$ $\exists \{\hat{x}_1, \ldots, \hat{x}_I\} : \hat{x}_{i+1} - \hat{x}_i \in \{0,1,2\},$ $x_{ij} - \hat{x}_i = 0, \ y_{i,j+1} - y_{ij} \in \{0,1,2\}$	
P2DHMDM	Pseudo 2-Dimensional Hidden Markov Distortion Model (first-order), allow horizontal displacements in P2DHMM $x_{1j} = 1, x_{Ij} = X, y_{i1} = 1, y_{iJ} = Y,$ $\exists \{\hat{x}_1, \ldots, \hat{x}_I\} : \hat{x}_{i+1} - \hat{x}_i \in \{0,1,2\},$ $x_{ij} - \hat{x}_i \in \{-1,0,1\}, \ y_{i,j+1} - y_{ij} \in \{0,1,2\}$	
IDM	Image Distortion Model (zero-order), disregard relative displacements of neighboring pixels, restrict absolute displacement $x_{ij} \in \{1, \ldots, X\} \cap \{i' - w, \ldots, i' + w\}, i' = \left[i\frac{X}{I}\right],$ $y_{ij} \in \{1, \ldots, Y\} \cap \{j' - w, \ldots, j' + w\}, j' = \left[j\frac{Y}{J}\right],$ with warp range w and $[\cdot]$ the nearest integer function	

best investigate the effect of the matching techniques. Additionally, we use the nearest prototype decision rule in a key experiment, which can strongly reduce the computation time in classification.

In summary, we can conclude that a large variety of matching models for images have been discussed in the literature. However, in this paper, we address several questions, outlined in Section 1, that have so far remained unanswered. These concern the direct performance comparison of the models in image recognition tasks and their generalization ability, the necessity of complex matching restrictions that lead to computationally complex algorithms, and the use of image context at the pixel level.

3 FRAMEWORK FOR RECOGNITION USING NONLINEAR MATCHING

For our discussion of the nonlinear deformation models and the experiments we performed, we use the following framework: We denote the test image (or observation) by $A = \{a_{ij}\}$, where the pixel positions are indexed by $(i,j), i = 1, \ldots, I, j = 1, \ldots, J$, and (x,y) denotes the pixel positions within the reference image (or model) $B = \{b_{xy}\}, x = 1, \ldots, X, \ y = 1, \ldots, Y$. At each image position, we observe a vector of values $a_{ij}, b_{xy} \in \mathbb{R}^U, a_{ij} = (a_{ij}^1, \ldots, a_{ij}^U),$ $b_{xy} = (b_{xy}^1, \ldots, b_{xy}^U)$ that can represent gray values $(U=1)$, color values $(U=3)$, the vertical and horizontal image gradient $(U = 2)$, or a larger pixel context (e.g., $U = 18$ for 3×3 pixel contexts of the image gradients).

We consider image deformation mappings that map pixel positions of the test image onto pixel positions of the reference image and must fulfill certain constraints

$$\begin{aligned} \left(x_{11}^{IJ}, y_{11}^{IJ}\right) : \\ (i,j) \mapsto (x_{ij}, y_{ij}), \ \ i = 1, \ldots, I, j = 1, \ldots, J, \ \left(x_{11}^{IJ}, y_{11}^{IJ}\right) \in \mathcal{M}. \end{aligned}$$
(1)

The set $\mathcal{M}$ of possible image deformation mappings defines the model used. Each model will be discussed in more detail in the following sections. A brief informal summary of the models is given along with their formal definitions in Table 1.

The distortion models with their permitted deformation mappings differ in the way they treat the interdependence of local pixel displacements. When a pixel (i, j) is mapped onto a reference pixel (x_{ij}, y_{ij}), we can observe the difference of this mapping to the mapping of its neighbors $(i - 1, j)$ and $(i, j - 1)$ in the original image. For example, to ensure a continuous and monotonic mapping, we do not allow the difference in displacement of neighboring pixels to be negative (no crossings) nor greater than two pixels. We also restrict the maximum horizontal displacement difference of vertically neighboring pixels to at most one pixel. This restriction would also apply, respectively, to neighboring pixels. This approach
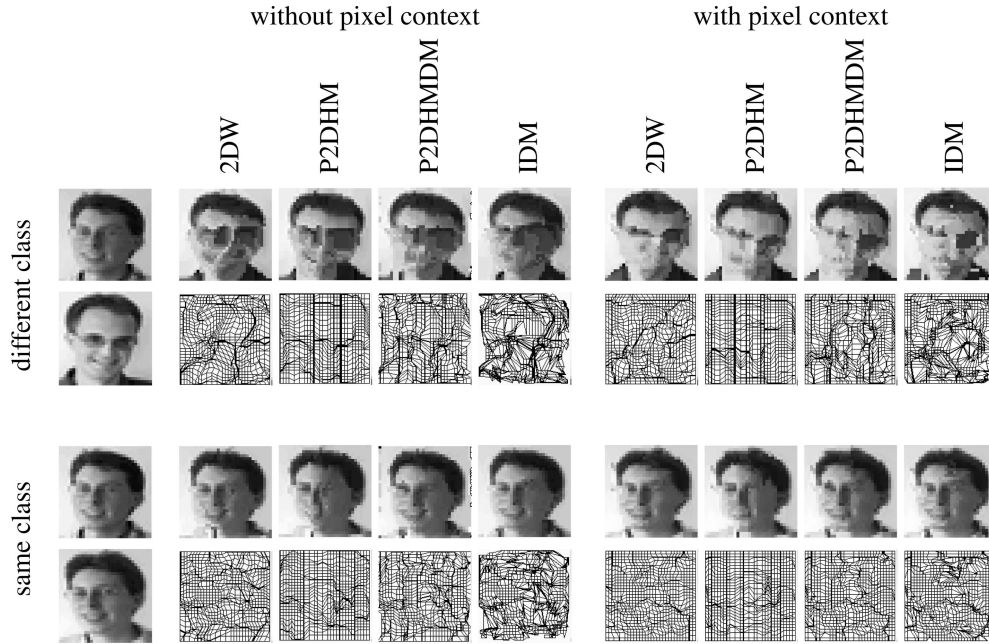
Fig. 1. Examples of nonlinear matching applied to face images (see text).

leads to the two-dimensional warping model, while the other models result if some of these restrictions are relaxed.

From the constraints presented in Table 1, we can observe that there are relative and absolute constraints: Relative constraints concern the relation between mappings of neighboring pixels, e.g., $x_{i+1,j} - x_{ij}$, while absolute constraints only look at the original position of the pixel in the image, e.g., possible values for x_{ij} only depend on i and j. The IDM only includes absolute constraints, which allows efficient minimization. Often, an absolute constraint like the one for the IDM is additionally imposed for the other models, e.g., the warp range is limited for the P2DHMM by adding to the constraints of the P2DHMM the constraints of the IDM. Note that these constants depend on the image and object size and resolution.

The constraints presented here are all hard constraints, i.e., it is either possible or not possible to have a certain mapping under a specific model. Although we will not discuss this possibility further, we could also use cost functions instead of these hard constraints, such that certain mappings are permitted, but only at a higher cost, where this cost is added to the distance value. (The hard constraints are a special case of the cost functions, where the cost takes only the values 0 and ∞.) Cost functions also allow us to additionally penalize some permitted mappings. For example, $x_{i,j+1} - x_{ij} = 0$ could be assigned cost zero and $x_{i,j+1} - x_{ij} > 0$ could involve higher costs, thus penalizing mappings that deviate considerably from the linear mapping. Some cost functions of this type are easily integrated into the algorithms and do not change the runtime significantly, while others may require algorithmic changes. Cost functions of the first kind have been evaluated in detail in informal experiments, but, in exchange for the number of additional parameters that need to be tuned, no significant improvements could be obtained.

Figs. 1 and 2 show sample images that result from a matching using the different distortion models discussed in this paper. We use face images in Fig. 1 because they allow a very intuitive understanding of dissimilarity and images of handwritten digits in Fig. 2 as they are used in the experiments. In both figures, the first column shows the test and reference images. The rest of the upper row shows the transformed reference images, using the respective models, that best match the test image. The lower row shows the respective displacement grids generated to obtain the transformed images. The two first rows show images of "different classes," while the other two rows show images of the "same class." The examples on the left were created using only the image gray values. We observe that, for the same class, the models with more restrictions produce inferior matches to the models with less restrictive matching constraints. We are interested in discriminating between different classes, therefore our goal is to have good matchings (small distances between the test image and the transformed reference image) for images of the same class, but matchings with large distances for images of different classes. This is obviously not achieved by the models with fewer restrictions, e.g., the IDM, when applied using only the gray values. The second four examples show the results for the case using local context for matching as detailed below. Note that the matchings for the same class remain very accurate, while the matching for the different classes are visually not as good as before. This is especially true for the models with fewer constraints, such as the IDM. Note also that the displacement grid is more homogeneous for the matchings of the same class.

Decision Rule. In this paper, the main emphasis lies on the effect of the different distance functions resulting from the various deformation models. We therefore choose a simple decision rule for the recognition process that does not introduce any new parameters that subsequently need to be adjusted. Assuming a reference data set of images $B_{k1}, \ldots, B_{kN_k}$ for classes $k = 1, \ldots, K$, we use the nearest neighbor decision rule, which is known to yield good results in various applications:

$$A \longmapsto \hat{k}(A) = \arg\min_k \{ \min_{n=1,\ldots,N_k} d(A, B_{kn}) \}. \tag{2}$$
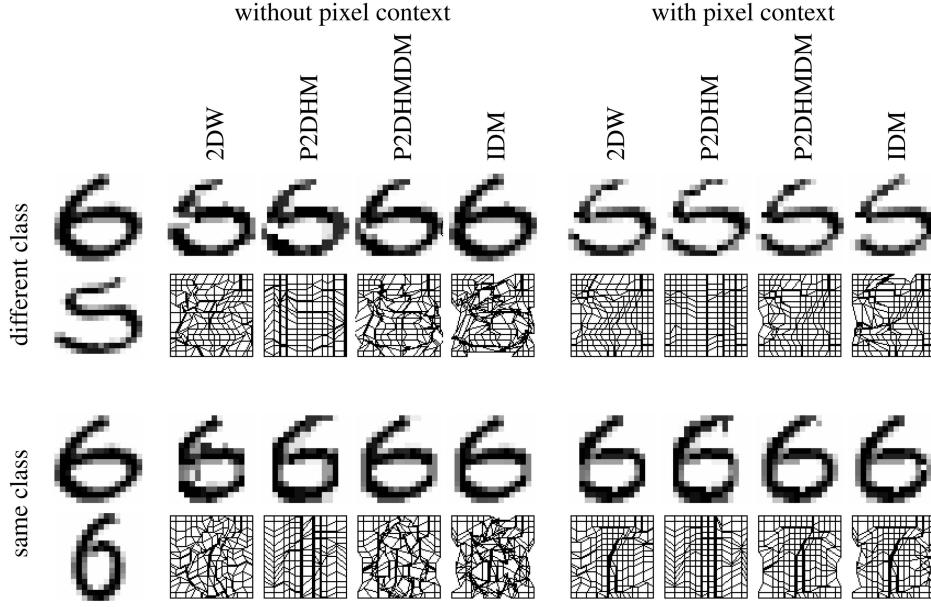
Fig. 2. Examples of nonlinear matching applied to USPS digit images (see text).

The distance function used within the decision rule has a special structure, it results from a minimization over all permitted deformation mappings for the model used

$$d(A, B) = \min_{(x_{11}^{IJ}, y_{11}^{IJ}) \in \mathcal{M}} \left\{ d'(A, B_{(x_{11}^{IJ}, y_{11}^{IJ})}) \right\}. \qquad (3)$$

Here, the minimization process is of varying computational complexity depending on the model used as discussed below. Note that we use an asymmetric distance function here because we want to find the reference that best *explains* the observed image, similar to the use of HMMs in other classification tasks. The simpler distance function d' used within the minimization is chosen to be the squared euclidean distance over all vector components of all pixels that are compared:

$$d'\left(A, B_{(x_{11}^{IJ}, y_{11}^{IJ})}\right) = \sum_{i,j} \sum_u \left(a_{ij}^u - b_{x_{ij}y_{ij}}^u\right)^2 = \sum_{i,j} \|a_{ij} - b_{x_{ij}y_{ij}}\|^2. \qquad (4)$$

To better analyze the minimizations for the different models, we can rewrite the distance function by factoring out the dependencies of $(x_{11}^{IJ}, y_{11}^{IJ}) \in \mathcal{M}$ by letting each choice at a pixel (i, j) depend only on choices made for pixels before reaching this pixel

$$d(A, B) = \min_{(x_{11}^{IJ}, y_{11}^{IJ})} \sum_{i,j} f(A, B, x_{ij}, y_{ij}, x_{11}^{i,j-1}, y_{11}^{i,j-1})$$

$$f(A, B, x_{ij}, y_{ij}, x_{11}^{i,j-1}, y_{11}^{i,j-1}) =$$
$$\begin{cases} \|a_{ij} - b_{x_{ij}y_{ij}}\|^2 & \text{if } x_{11}^{i,j-1}, y_{11}^{i,j-1} \text{ extensible by } x_{ij}, y_{ij} \text{ in } \mathcal{M} \\ \infty & \text{otherwise.} \end{cases}$$
$$(5)$$

Although this formulation may look more complicated than necessary, we can now immediately analyze the following cases:

$$\text{2DW}: d(A, B)$$
$$= \min_{(x_{11}^{IJ}, y_{11}^{IJ})} \sum_{i,j} f(A, B, x_{ij}, y_{ij}, x_{i,j-1}, y_{i,j-1}, x_{i-1,j}, y_{i-1,j}),$$
$$(6)$$

$$\text{P2DHMM}: d(A, B)$$
$$= \min_{(x_{11}^{IJ}, y_{11}^{IJ})} \sum_{i,j} f(A, B, x_{ij}, y_{ij}, x_{i,j-1}, y_{i,j-1}), \qquad (7)$$

$$\text{IDM}: d(A, B) = \sum_{i,j} \min_{(x_{ij}, y_{ij})} f(A, B, x_{ij}, y_{ij}) \qquad (8)$$

(with the canonical predecessor for the limit cases, e.g., $j = 1$). We see that, for the zero-order IDM, f is independent of the history and, therefore, the minimization and the summation can be interchanged, leading to a computationally simple minimization problem. For the first-order P2DHMM, the dependence is restricted to immediate successors. This means that the minimization can be done using dynamic programming with recombination over only one variable and, thus, an efficient algorithm exists. (The P2DHMDM case is similar but requires more tedious notation.) Only for the second-order 2DW model does the dynamic programming algorithm need to recombine over J previous variables and, therefore, the algorithm is exponential in the image size; in fact, the problem can be shown to be NP-complete [27].

Feature extraction. An analysis of matching results when using only pixel gray values shows that often unwanted deformations are observed that allow good matchings for images of different classes. This behavior can be restricted by including the local image context of each pixel in an appropriate way, which is confirmed by experimental results: In experiments on the USPS corpus only, a comparatively small improvement from 5.6 percent (no matching) to 4.0 percent was possible when using the pixel values directly. This error rate can be approximately halved when using pixel context information.
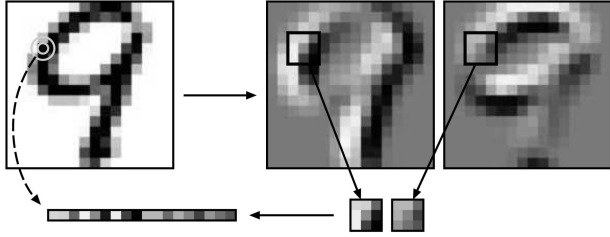
Fig. 3. Local context extraction, 3×3 subimages of gradients.

One straightforward 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. Experiments showed that best results are obtained when using only gradient information and disregarding the image brightness values completely [24].

The use of gradient information is a fairly standard approach in many image processing and recognition contexts. It is, for example, also used by [36] (histograms of gradient information around interest points) and by [2] (histograms of contour point occurrence around selected contour points). In this paper, we show how it can be used to achieve excellent recognition results by counteracting the effect of using fewer restrictions in the matching process for recognition.

Of course, when using the first derivative, it is natural to ask if the second derivative could lead to similar improvements. Unfortunately, the additional use of the second derivative led to only small improvements in some of the experiments performed and to degraded performance in some other cases. Therefore, we chose not to use higher order derivatives.

A second way to include the local image context is to use local subimages that are extracted around the pixel concerned, e.g., of size 3×3 pixels. Naturally, the size of the optimal context depends on the resolution of the objects in the images. Thus, this size should be determined for each task individually. For handwritten character recognition with image sizes between 16×16 and 28×28 pixels, we observed that a local context of 3×3 pixels leads to the best results, tested for the USPS data set. The performance generalized well for the other digit recognition tasks and medical images. In the experiments, the medical images were scaled down to a common height of 32 pixels from varying, much larger sizes using standard averaging. This reduction was used to reduce the runtime and because we observed that this size was a good comprise for the task of recognition, containing enough detail to allow a clear distinction of the classes, while reducing the extraneous details that are counterproductive in classification.

The contexts can be extracted from the gray values in the image, thus leading to a vector of dimension $U = 9$. This already leads to improvements in recognition. However, we can combine the two methods and extract the contexts from the gradient images, which changes the value of an image pixel to a vector of dimension $U = 18$. This combination leads to consistently better results and was therefore adopted in all experiments presented in the following. Fig. 3 illustrates the procedure of context extraction.

4 SECOND-ORDER: TWO-DIMENSIONAL MODEL

We begin the discussion of the two-dimensional matching algorithms with the true two-dimensional model. The two-dimensional HMM or two-dimensional warping (2DW) is an extension to two dimensions of the (0, 1, 2)-HMM that is frequently used in applications such as speech recognition and the recognition of continuous handwriting. The model ensures monotonicity (no backward steps) and continuity (no large jumps) of the displacement grid. Note that the term HMM is used here for a distance model as is done in large parts of the relevant literature, although no probabilistic view is taken (no parameter estimation, no estimation of transition probabilities). In contrast to the one-dimensional case, which allows polynomial solutions, the minimization of this true 2D model is NP-complete [27]. Therefore, approximation algorithms are used, for example, dynamic programming with beam search [52] or simulated annealing. For a detailed discussion of the model and the dynamic programming algorithm that can be used for the minimization, we refer to the works of Uchida and Sakoe [52], [54].

The 2DW model results from imposing the following constraints on the image deformation mappings $(x_{11}^{IJ}, y_{11}^{IJ})$ permitted within the matching for recognition (3) (see Table 1). We first impose the border constraint that image borders should be matched on image borders: $x_{1j} = 1, x_{Ij} = X$, $y_{i1} = 1, y_{iJ} = Y$. This constraint can be most easily relaxed by padding the images with background pixels. Second, we require that horizontally adjacent pixels should not be matched onto pixels that deviate from the relative position in the original image by more than one pixel. To do so, we need two constraints, one in the horizontal direction, $x_{i+1,j} - x_{ij} \in \{0, 1, 2\}$, and one in the vertical direction, $x_{i,j+1} - x_{ij} \in \{-1, 0, 1\}$. The same constraints are imposed for vertically adjacent pixels: $y_{i,j+1} - y_{ij} \in \{0, 1, 2\}, y_{i+1,j} - y_{ij} \in \{-1, 0, 1\}$.

There are several possible algorithms to determine an approximation of the best matching under the 2DW model, as discussed above. In this section, we briefly describe the algorithm based on dynamic programming, as introduced by [52]. Theoretically, this algorithm allows us to obtain the optimal solution, but only at runtimes exponential in the minimum of the image widths. To obtain a solution in a feasible amount of time, we use the concept of beam search. More specifically, in each stage (i, j) (corresponding to the pixel positions) of the algorithm, we define a threshold value by increasing the score of the best hypothesis by a fixed value, then we disregard all hypotheses with scores higher than the threshold value. Furthermore, we limit the maximum number of hypotheses to a fixed value (500 in the experiments). Algorithm 1 gives an outline of the algorithm used to calculate this distance. There, $\mathcal{M}'(i, j)$ denotes the set of all possible submappings for the last J pixels up to (i, j) in the order described above that can be extended to a permitted mapping in $\mathcal{M}$. The notation $pre(s) \subset \mathcal{M}'(i, j - 1)$ denotes all possible predecessors of such a submapping s that are compatible with s in the sense that they agree in their mappings for all shared pixel positions and that the union of their mappings can be extended to a permitted mapping in $\mathcal{M}$. Note that we do not elaborate on all special cases here, e.g., for $i = 1$ or $j = 1$, the recursion obviously must take a different form.

The size of $\mathcal{M}'(i, j)$ is exponential in J, which makes the algorithm difficult to apply without beam search, even for only moderately large J.

Algorithm 1 2DW-distance; input: test image A,
reference image B;
for $i = 1$ to I
　　for $j = 1$ to J
　　　　for all separating submappings $s \in \mathcal{M}'(i,j)$
　　　　　　//(note: $|\mathcal{M}'(i,j)|$ exponential in J)
　　　　　　$Q(i,j,s) = \|a_{ij} - b_{s(i,j)}\|^2 + \min_{s' \in pre(s)} Q(i, j-1, s')$
output: $\min_{s \in \mathcal{M}'(I,J)} Q(I,J,s)$

Other possible strategies to determine an approximation of the best matching include: simulated annealing, which is a common strategy to find approximate solutions to hard optimization problems, turbo-decoding, i.e., iterative matching with relaxed horizontal and vertical constraints [42], and piecewise linear matching [43].

5　FIRST-ORDER: PSEUDO-TWO-DIMENSIONAL MODELS

The complexity of the true two-dimensional matching algorithms—even when using approximations like beam search or simulated annealing—is very high. To be able to apply deformation models for real-world tasks, we must therefore consider models of a lesser order.

Pseudo-two-dimensional hidden Markov model. To proceed from two-dimensional models to models of lesser order, we relax some of the constraints of the true two-dimensional case. Usually, this is done by allowing the columns of an image to be matched onto the reference image independently, which leads to the so-called pseudo-two-dimensional hidden Markov model (P2DHMM) [30]. The P2DHMM is obtained from the 2DW model by neglecting relative displacement in the vertical image direction between pixels of neighboring image columns and mapping all pixels from one column onto the same target column. We impose the same border constraints as for the 2DW model and require that a mapping of image columns $\{\hat{x}_1, \ldots, \hat{x}_I\}$ exists for which $\hat{x}_{i+1} - \hat{x}_i \in \{0, 1, 2\}$ and each $x_{ij} = \hat{x}_i$, i.e., the horizontal displacement is the same for all pixels of one column. For the vertical displacements, we keep the constraints $y_{i,j+1} - y_{ij} \in \{0, 1, 2\}$ (see Table 1).

The assignment of complete columns onto other columns has two consequences: First, only complete columns of the reference image can be skipped in the matching. Second, dependencies between the vertical displacements of the pixels in neighboring columns are ignored. The first of these limitations is avoided to some degree by allowing additional deviations from the column assignment as described below. The second limitation cannot be overcome easily without arriving at a second-order problem again.

The main direction of the model is usually chosen left-to-right. This is the canonical form for western text, but is an arbitrary decision for general objects or for isolated characters. The algorithm can be applied similarly with the main direction of the model being top-to-bottom.

Pseudo-two-dimensional hidden Markov distortion model. To relax the constraint that complete columns of the images must always be matched, we can allow additional distortions from the columns that are matched by an additional pixel within the P2DHMM model. These distortions are modeled to be independent of each other and we call the resulting model pseudo-two-dimensional hidden Markov distortion model (P2DHMDM) [24], [23]. This abbreviation is

rather lengthy but clearly shows the relationship to the P2DHMM and also to the image distortion model (IDM) as discussed below. Formally, we only need to relax the constraint $x_{ij} - \hat{x}_i = 0$ in the P2DHMM to $x_{ij} - \hat{x}_i \in \{-1, 0, 1\}$ (see Table 1). The mapping of one column of the test image is determined by dynamic programming similar to the case of the P2DHMM, only now a deviation from the target column is permitted, which is at most one pixel for this example and also for the experiments reported here. This additional flexibility within the mapping of the columns leads to a slightly higher computational complexity.

The algorithms for both the P2DHMM and the P2DHMDM are given in Algorithm 2.

Algorithm 2 P2DHMM-distance; input: test image A,
reference image B;
for $i = 1$ to I
　　for $x = 1$ to X
　　　　for $j = 1$ to J
　　　　　　for $y = 1$ to Y
　　　　　　　　$Q(i,j,x,y) = \|a_{ij} - b_{xy}\|^2$
　　　　　　　　　　$+ \min_{y' \in \{y, y-1, y-2\}} Q(i, j-1, x, y')$

　　　$Q'(i,j) = Q(i,x,J,Y) + \min_{x' \in \{x, x-1, x-2\}} Q'(i-1, x')$
output: $Q'(I,X)$
for the **P2DHMDM**, replace "$\|a_{ij} - b_{xy}\|^2$"
with "$\min_{x' \in \{x-1, x, x+1\}} \|a_{ij} - b_{x'y}\|^2$"

6　ZERO-ORDER: THE IMAGE DISTORTION MODEL

If we further relax the constraints on the image mapping functions and impose only absolute constraints, we arrive at a zero-order model of image variability. Its advantage is that the minimization process is computationally much simpler, but high recognition performance can be achieved when using an appropriate local image context representation. Formally, we require that each test image pixel is mapped to a pixel within the reference image not more than w pixels from the place it would take in a linear matching: $x_{ij} \in \{1, \ldots, X\} \cap \{i' - w, \ldots, i' + w\}$, $i' = [i\frac{X}{I}]$, $y_{ij} \in \{1, \ldots, Y\} \cap \{j' - w, \ldots, j' + w\}$, $j' = [j\frac{Y}{J}]$.

An informal description of the model is the following: For each pixel in the test image, determine the best matching pixel (possibly including its context) within a region of size $w \times w$ defined around the corresponding position in the reference image and use this match. The formal constraints are given in Table 1. Due to its simplicity and efficiency, this model has been described independently in the literature several times with differing names, see [47], the references therein, and the references in Section 2. However, this model has not usually incorporated pixel contexts in the past, which is the key to achieving low recognition error rates. The model is called the image distortion model (IDM) here.

Since the dependencies between pixel displacements are neglected, the minimization process for the IDM involves only a local optimization for each pixel position in the test image and, thus, is computationally inexpensive. In comparison with the euclidean distance, the required computation time is increased by a factor of approximately $(2w + 1)^2$,
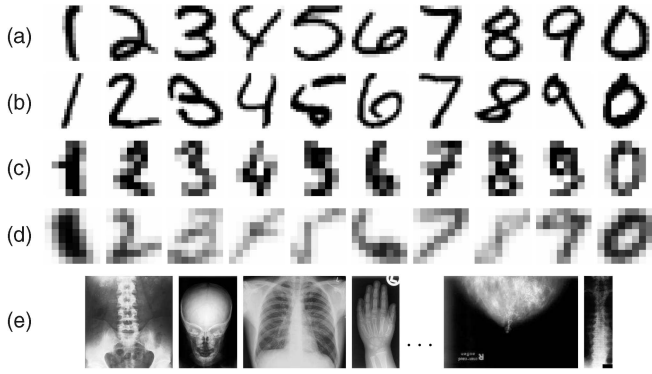
(a)
(b)
(c)
(d)
(e)

Fig. 4. Examples (a) USPS, (b) MNIST, (c) UCI, (d) MCEDAR, and (e) IRMA/ImageCLEF.

where w denotes the warp range, i.e., the maximum permitted absolute displacement of any one pixel.

In earlier experiments, we did not include local pixel context for the matching in the IDM [21]. Without local pixel content the IDM already provided improved results for radiograph recognition, but was not able to improve the results for handwritten digit recognition. The inclusion of local context led to a noticeable improvement of the performance in both applications and we could generalize the improvements across different data sets [23], [24], [58].

The IDM is a very natural approach. Nevertheless, it is an effective means of compensating for small local image variations and the intuitiveness of the model may also be seen as an advantage: It is almost as easily implemented as the euclidean distance. An outline of the algorithm is given in Algorithm 3.

Algorithm 3 IDM-distance; input: test image A,
reference image B;
for $i = 1$ to I
    for $j = 1$ to J
        $i' = \left[i\frac{X}{I}\right],\ j' = \left[j\frac{Y}{J}\right]$
        $s = s + \min\limits_{\substack{x \in \{1,...,X\} \cap \{i'-w,...,i'+w\} \\ y \in \{1,...,Y\} \cap \{j'-w,...,j'+w\}}} \|a_{ij} - b_{xy}\|^2$
output: s

Zero-order matching using the Hungarian algorithm. The task of finding a matching between pixels in an observed image and those in a reference image is common to all deformation models discussed in this paper. The term "matching" is a well-known expression in graph theory, where it refers to a selection of edges in a (bipartite) graph. Such a matching can be constructed using, for example, the Hungarian Algorithm [29, p. 74ff], which has been used before in the context of image recognition [2].

Using the Hungarian algorithm, it is possible to impose the global constraint that each pixel of both compared images must be matched at least once. This method is called the Hungarian Distortion Model (HDM) [22] and it takes into account larger parts of the reference images and, as a result, more homogeneous displacement fields are obtained. Nonetheless, as displacements of neighboring pixels are not accounted for, the HDM is a zero-order model, as is the IDM on which it is based.

The complexity of the Hungarian algorithm is cubic in the number of vertices, which results in a problematic

TABLE 2
Corpus and Image Sizes

name	size	K	#train	#test
USPS	16×16	10	7 291	2 007
MNIST	28×28	10	60 000	10 000
UCI	8×8	10	3 823	1 797
MCEDAR	8×8	10	11 000	2 711
IRMA	varying	57	9 000	1 000

TABLE 3
Summary of Results for Handwritten Digit Recognition (ER [%])

database	2DW	P2DHMM	P2DHMDM	IDM
USPS	2.7	2.5	1.9	2.5
MNIST			0.52	0.54
UCI		1.1	0.8	0.8
MCEDAR			3.3	3.5
time [ms]	590.9	11.9	44.6	0.5

Timing for one USPS distance computation on a 3 GHz PC.

runtime in the order of $O(\max((IJ)^3, (XY)^3))$. For example, for the USPS task, this results in the duration of 42 ms per image comparison on a 3 GHz PC.

7 RESULTS

In this section, we discuss the results obtained using the previously described models on the databases described in the following. We start with the results obtained by the two-dimensional model, decrease complexity to the first and zeroth order model, and, finally, present results for the training of prototypes using the discussed models. A summary is given in Table 3.

Databases. Fig. 4 shows example images and Table 2 shows an overview of the statistics for the databases used in the experiments. A detailed description of each database is presented in the Appendix.

USPS: The US Postal Service task is still one of the most widely used reference data sets for handwritten character recognition and allows fast experiments due to its small size. The test set contains a large amount of image variability and is considered to be a "hard" recognition task. Good error rates are in the range of 2-3 percent.

MNIST: The modified NIST database can be considered the standard benchmark for handwritten character recognition at the moment. A large number of reference results are available. The MNIST data set is larger in size than the USPS data set and contains less variability in the test set. Here, good error rates are in the range of 0.5-1.0 percent.

UCI & MCEDAR: Both the digit database from the University of California, Irvine (UCI) Repository of Machine Learning Databases and the Modified CEDAR (Center of Excellence for Document Analysis and Recognition) database contain images of much lower resolution than USPS and MNIST. For these data sets, only a few reference results are available. They were included in the experiments to show that the presented methods generalize well

and were not optimized to specific data sets. For the experiments, we scale the images up to 16×16 pixels using spline interpolation. Good error rates here lie in the range of 1-2 percent and 3.5-4.5 percent, respectively.

IRMA: The IRMA/ImageCLEF 2005 database contains medical images from daily hospital routine (IRMA = Image Retrieval in Medical Applications, CLEF = Cross Language Evaluation Forum). The database was used as a part of the 2005 ImageCLEF workshop for the evaluation of image retrieval systems in which 12 groups submitted results that can serve as a basis of comparison. The task here is to determine the correct category among 57 choices and good error rates are in the range of 12-15 percent.

Classifier settings and baseline results. In all reported experiments, the 3-NN classifier was used for the decision making as it often performs better on the average than the 1-NN classifier. In all experiments (except when using single prototypes), all of the training images in a data set were used as references. For some experiments with very time consuming distance calculations, the final distance was only computed for the, e.g., 500 closest references based on the euclidean distance. We compared the results of the classifier using no distortion on the USPS data set and observed that the error rate of 5.5 percent using gray values increased to 6.4 percent using the 18-dimensional context feature vector. This difference is due to the attenuation of the influence of absolute gray value differences when using derivatives, which has a negative effect on the classification performance for this task when not allowing matching. This highlights that it is the *combination* of local context and distortion model that leads to good performance.

Recognition results using the 2DW. Due to its exceptionally high computational complexity, we performed only a small number of experiments using the true two-dimensional deformation model. All of these experiments used the USPS data because it is comparatively small both in the number of samples and in the size of the images. The best result we obtained was an error rate of 2.7 percent. It is possible that better results could be obtained using greater search effort (i.e., larger beam sizes). But, observing the very good error rates, at much lower computational complexity, of the simple models described in the following sections, it seems unlikely that results better than these will be observed, even with much additional search effort.

Recognition results using the P2DHMM and the P2DHMDM. Using the P2DHMDM, we were able to improve the results for different recognition tasks, both with respect to the P2DHMM and in general.

On the USPS task, the error rate of the P2DHMM is 2.5 percent, which can be reduced to 1.9 percent using the additional flexibility of the P2DHMDM with local pixel contexts [24]. On the USPS database, the P2DHMDM performed better than the other models and gave the best result published so far. On the UCI optical digits corpus, the P2DHMM yielded an error rate of 1.1 percent, while the P2DHMDM yielded 0.8 percent, the best published error rate. After observing these results, we generally preferred the P2DHMDM to the P2DHMM and, therefore, only present results for the former. On the MCEDAR task, the P2DHMDM again achieved the best published error rate of 3.3 percent. On the MNIST data, the excellent error rate of 0.52 percent could be achieved using the P2DHMDM.

Recognition results using the IDM. On the USPS task, the IDM achieves a very good error rate of only 2.5 percent, which is surprising for such a simple model. (Note that a 1-NN in this case performs slightly better at 2.4 percent error rate [24].) For example, to achieve the same result, a kernel density classifier using tangent distance has to include two-sided tangent distance and nine-fold virtual training and test data [21]. A warp range $w = 2$ is used for the IDM.

On the MNIST task, the use of the IDM resulted in an error rate of 0.54 percent, which is not significantly higher than the result of 0.52 percent using the P2DHMDM. This result also compares very well to the results of other approaches on the same data, although slightly better results are reported. The statistically insignificant improvement of the P2DHMDM probably does not justify the increased complexity of the P2DHMDM over the IDM in this case.

On the UCI optical digits corpus, the IDM performed as well as the more complex P2DHMDM, both achieved an error rate of 0.8 percent, which is the best error rate known. On the MCEDAR task the IDM achieves an error rate of 3.5 percent, which is only slightly higher than the best published error rate of 3.3 percent obtained by the P2DHMDM.

In the automatic annotation task of the 2005 ImageCLEF evaluation of content-based medical image retrieval using the IRMA data, the error rate of 12.6 percent obtained by the IDM was the best among 42 results submitted. The second best result uses the image distortion model along with the normalized cross covariance of gray values and Tamura texture features. In comparison, the baseline error rate obtained by a 1-nearest neighbor classifier using 32×32 images is 36.8 percent. The average error rate of all submissions was 32.7 percent and the median was 22.3 percent. Also, in the medical retrieval task of the 2004 ImageCLEF evaluation, the IDM was used in the best submission in the category "visual information only, no user interaction."

Recognition results using the HDM. Using the HDM on the USPS database, an error rate of 2.2 percent could be reached, which is an improvement over the 2.4 percent error rate achieved using the IDM alone. However, this does not justify the use of the greatly increased computational effort.

Recognition results using trained prototypes. We performed some experiments for training of prototypes with the presented matching models on the USPS corpus to investigate the performance of the models with fewer prototypes. Using the training data, a reference model for each of the 10 classes was estimated as described in the following. This model was then used as a single reference for that class for testing.

The prototypes are initialized using the mean images of each class (see Fig. 5, upper row). Then, each training image is matched to the prototype. Averaging the pixel values that the prototype pixels are matched to, a new set of prototypes is obtained that better represents the training data and takes into account the variability of the training images. This procedure is iterated until the prototype images do not change any more. Fig. 5 shows the mean images for all 10 classes and the prototypes learned using the P2DHMDM.

Fig. 5 also contains error rates obtained using prototypes trained using the P2DHMDM. These show that the P2DHMDM performs significantly better using the learned prototypes. Interestingly, using only one prototype per

means (no matching)

learned prototypes

error rates [%]

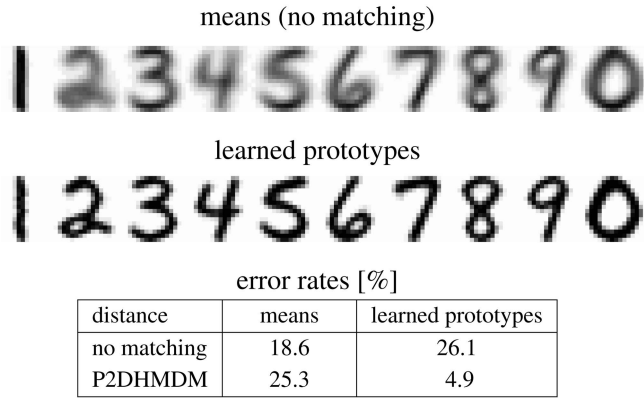| distance	means	learned prototypes
no matching	18.6	26.1
P2DHMDM	25.3	4.9

Fig. 5. Prototypes for the USPS training set using no matching (upper row) and using image matching (lower row) and the respective error rates for the USPS data.

class, the error rate is as low as 4.9 percent. The learned prototypes appear much less blurred than the mean images as the image variability is compensated for by the nonlinear deformation model. On the other hand, the corresponding mean images perform better if no deformation is used in the classifier, as could be expected because they represent the maximum likelihood estimate for that case.

Other results that these error rates can be compared to (i.e., that also use a single prototype per class) are briefly described in the following: Using the empirical means and an estimated, 14-dimensional tangent vector subspace, as in [25], the error rate obtained was 5.0 percent. Hastie and Simard [18] report an error rate of 4.1 percent using a 12-dimensional trained subspace approach that also changes the means. We can thus conclude that the use of the described models of image variability also gives state-of-the-art results for the use of single prototypes per class.

8 CONCLUSION

We discussed several nonlinear models of image variability that can be used as distance measures in appearance-based classifiers. From experiments on several real-world image recognition tasks, as summarized in Table 3, we can conclude the following: The simplest model—the image distortion model—used with pixel level features that describe the local image context represents the best compromise between computational complexity and recognition accuracy, while the more complex P2DHMDM leads to slightly better results on average and outperforms the conventional P2DHMM. The most complex model, the 2DW, performed worse or could not be evaluated due to the high-computational demand.

We can draw the following conclusion from the analysis of the experiments:

- The approach used of distortion modeling is appropriate for the classification tasks discussed, because it corresponds well to the deformations encountered in the images, including small affine transformations and local distortions. This can be seen not only by observing the low error rates, but also by regarding the very clear resulting prototypes as shown in Fig. 5.
- While the 2DW second-order model is conceptually very appealing because it takes into account the full 2D-constraints, the minimization under its constraints

is computationally very expensive, and approximations are necessary. This, in turn, leads to higher error rates in the experiments at still large runtimes.
- The first-order P2DHMM suffers somewhat from the very strict assignment of image columns to image columns which limits the range of transformations that can be modeled.
- These restrictions are alleviated for the zero-order IDM, which means, on the other hand, that more unwanted transformations can be modeled. By using the pixel context information, the unwanted transformations are made much more costly such that this efficient model achieves very good error rates.
- The best compromise between restricting the wanted transformations and allowing unwanted transformations seems to be the P2DHMDM with the use of local contexts. It combines positive effects of P2DHMM and IDM and leads to the overall best recognition performance at still acceptable runtimes.

With the discussed methods, we achieved very competitive results across four different handwritten digit recognition tasks, in particular an error rate of 0.52 percent on the MNIST task. We also showed that the same methods can be effectively used for the categorization of medical images. Although not discussed in this paper, we would like to mention that improvements in the domain of image sequence recognition (sign language word recognition, gesture recognition) can also be achieved by modeling the image variability with these deformation models [58].

We were able to show that the IDM with local gradient contexts leads to excellent results for the recognition of handwritten digits and medical images. On all data sets considered, state-of-the-art results were obtained using the efficient IDM, which can be described in a few statements and implemented in a few lines of code. To emphasize this point, we present the algorithm here, including the context extraction. Algorithm 4 was used for image matching for digit recognition and achieved an error rate of 0.54 percent on the MNIST data set.[1]

Algorithm 4 IDM-distance $(3 \times 3$-context); input:
test image A, reference image B;
$A^v = $ vertical Sobel (A), $A^h = $ horizontal Sobel (A)
$B^v = $ vertical Sobel (B), $B^h = $ horizontal Sobel (B)
for $i = 1$ to I
    for $j = 1$ to J
        $i' = \left[i\frac{X}{I} \right]$, $j' = \left[j\frac{Y}{J} \right]$

$$s = s + \min_{\substack{x \in \{1,...,X\} \cap \{i'-w,...,i'+w\} \\ y \in \{1,...,Y\} \cap \{j'-w,...,j'+w\}}} \sum_{m=-1}^{1} \sum_{n=-1}^{1} (A^v_{i+n,j+m} - B^v_{x+n,y+m})^2$$
$$+ (A^h_{i+n,j+m} - B^h_{x+n,y+m})^2$$

output: s

This distance algorithm used within a 3-NN classifier combines several properties that make it an ideal baseline algorithm for image recognition tasks in the presence of image variability:

1. The software used is available at http://www-i6.informatik.rwth-aachen.de/~gollan/w2d.html.

TABLE 4
Overview of Published Error Rates for the USPS Task [%]

ref.	group	method	ER
[46]	AT&T	human performance	2.5
[12]	CENPARMI	human performance	1.5
–	–	Euclidean nearest neighbor	5.6
[50]	Microsoft	relevance vector machine	5.1
[32]	AT&T	neural net LeNet1	5.0
[4]	AT&T	neural net	4.2
[21]	RWTH-i6	kernel densities, virtual data	4.2
[44]	TU Berlin	support vector machine	4.0
[17]	LMB+RWTH-i6	support vector m. + tangent distance	3.6
[16]	LMB	support vector m. + invariant features	3.5
[9]	RWTH-i6	LDA, virt. data, Gauss. mix. dens.	3.4
[10]	AT&T	local learning	*3.3
[44]	TU Berlin	invariant support vector machine	3.0
[21]	RWTH-i6	two-sided tangent distance	3.0
[24]	RWTH-i6	3-NN, 2-D deformation model	2.7
[14]	AT&T	neural net + boosting	*2.6
[46]	AT&T	tangent distance	*2.5
[13]	CENPARMI	preprocessing, support vector machine	2.5
[21]	RWTH-i6	two-sided tangent distance, virtual data	2.4
[24]	RWTH-i6	1-NN, zero-order deformation model	2.4
[13]	CENPARMI	preprocessing, virtual support vector m.	2.2
[22]	RWTH-i6	deformation, Hungarian matching	2.2
[26]	ITI+RWTH-i6	tangent distance + local patches	2.0
[24]	RWTH-i6	3-NN, pseudo 2-D deformation model	1.9

Error rates using the USPS+ training set are indicated by *.

- it yields very good results across different tasks,
- it is easy to implement,
- it is computationally efficient, and
- it has only a few free parameters.

APPENDIX

In the appendix, we describe each of the data sets used in more detail along with available reference results.

A.1 The US Postal Service Task

The well-known United States Postal Service Handwritten Digit Database (USPS) consists of isolated and normalized images of handwritten digits taken from US mail envelopes scaled to 16×16 pixels. The database contains a separate training and test set, with 7,291 and 2,007 images, respectively. Different versions of this database exist. The experiments performed here are based on the data as available via FTP from the Max Planck Institute Tübingen.[2] A slightly different version of the USPS data exists, sometimes called USPS+ [10]. Here, 2,549 additional images of machine printed digits were added to the training set [31].

One disadvantage of the USPS corpus is that no development test set exists, resulting in the possible underestimation of error rates for all of the reported results. Note that this disadvantage holds for almost all data sets available for image object recognition.

One advantage of the USPS task is the availability of many recognition results reported by international research groups, allowing a meaningful comparison. Results for

TABLE 5
Error Rates for the MNIST Task [%]

ref.	group	method	ER
[46]	AT&T	human performance	0.2
–	–	Euclidean nearest neighbor	3.5
[37]	U Lige	decision trees + sub-windows	2.63
[32]	AT&T	deslant, Euclidean 3-NN	2.4
[38]	UC London	products of experts	1.7
[40]	U Québec	hyperplanes + support vector m.	1.5
[44]	TU Berlin	support vector machine	1.4
[4]	AT&T	neural net LeNet4	1.1
[46]	AT&T	tangent distance	1.1
[21]	RWTH-i6	two-sided tangent d., virt. data	1.0
[32]	AT&T	distortions, boosted LeNet4	0.7
[48]	U Singapore	bio-inspired features + SVM	0.72
[10]	Caltech+MPI	virtual SVM (box jitter)	0.68
[2]	Berkeley	shape context matching	0.63
[49]	U Singapore	deslant, biology-inspired features	0.59
[10]	Caltech+MPI	virtual SVM (box jitter + shift)	0.56
[24]	RWTH-i6	deformation model (IDM)	0.54
	RWTH-i6	deformation model (P2DHMDM)	0.52
[35]	Hitachi	preprocessing, support vector m.	0.42
[45]	Microsoft	neural net + virtual data	0.42
[11]	CENPARMI	virtual SVM (smoothing + shift)	0.38

different algorithms are listed in Table 4. The USPS data set continues to be used in a number of recent publications.

A.2 The MNIST Task

The modified NIST (National Institute of Standards and Technology) handwritten digit database (MNIST, [32]) is very similar to the USPS database in its structure. The main differences are that the images are not normalized and that the corpus is much larger. It contains 60,000 images in the training set and 10,000 patterns in the test set of size 28×28 pixels. The data set is available online.[3] This data set is generally considered to be an easier recognition task than the USPS data for two reasons. First, the human error rate is estimated to be only 0.2 percent, although it has not been determined for the whole test set [46]. Second, the (almost 10 times) larger training set allows machine learning algorithms to generalize better. Table 5 gives an overview of the error rates reported in other publications for the MNIST data.

Note that the methods used in this paper were not optimized for the MNIST task. The same parameters that proved to work well on the USPS data were chosen without further tuning.

A.3 The UCI Task

The data of the UCI task was obtained from the University of California, Irvine (UCI) Repository of Machine Learning Databases[4] [39]. The data set contains handwritten digits of size 8×8 pixels with 17 gray levels. It is separated into a training set of 3,823 images and a test set comprising 1,797 images. Its construction is described in more detail in [1].

---

2. ftp://ftp.kyb.tuebingen.mpg.de/pub/bs/data.

3. http://www.research.att.com/~yann/ocr/mnist/.
4. ftp://ftp.ics.uci.edu/pub/machine-learning-databases/~optdigits.

TABLE 6
Results for the UCI Task, Error Rates [%]

ref.	group	method	ER
		Euclidean 1-NN	2.0
[1]	Bogazici U, Istanbul	cascading classifier	4.7
[28]	Pohang U, S. Korea	PCA 12 comp.	2.2
[28]	Pohang U, S. Korea	2 PCA mixtures, 12 comp.	1.5
[24]	RWTH-i6	nonlinear deformation model	0.8

TABLE 7
Error Rates [%] for the MCEDAR Data Set

ref.	group	method	ER
		Euclidean 1-NN	5.7
[19]	U Toronto	PCA	4.9
[3]	Microsoft	Bayesian PCA	4.8
[19]	U Toronto	factor analysis	4.7
[51]	Microsoft	probabilistic PCA	4.6
[24]	RWTH-i6	3-NN, zero-order distortion model	3.5
[24]	RWTH-i6	3-NN, pseudo 2D distortion model	3.3

TABLE 8
Overview of Results Achieved in the 2005 ImageCLEF
"Automatic Annotation Task" Evaluation

group	method or label	ER[%]
U Montreal	texture features	73.3
NCTU	SVM to learn image characteristics	61.5
U Montreal	feature combinations	55.7
CEA	quantified colors, 9 nearest neighbor	46.0
Concordia U	SVM, mixed image feature vectors	43.3
Mt.Holyoke	Gabor Energy features	40.3
Mt.Holyoke	Gabor Energy features	37.8
CEA	image projection, 3-NN	36.9
–	32×32, 1-NN	36.8
NCTU	SVM to learn image characteristics	24.7
NTU	gray value blocks	22.5
U Madrid	GIFT + nearest neighbor	22.3
U Geneva	medGIFT retrieval	22.1
Infocomm	texture features + SVM	21.7
NTU	gray value block + nearest neighbor	21.7
U Madrid	GIFT + decision table	21.4
Infocomm	texture features + SVM	20.6
U Geneva	medGIFT retrieval	20.6
U Liege	patches, trees	14.7
RWTH-MI	IDM + Correlation + Tamura	14.6
U Liege	patches, boosting	14.1
RWTH-i6	patches, discriminative training	13.9
RWTH-MI	IDM + Correlation + Tamura	13.3
RWTH-i6	IDM, local context	12.6

Fig. 2 shows some example images from the UCI task and Table 6 shows a summary of the available results.

A.4 The MCEDAR Task

The modified CEDAR (MCEDAR) data set is based on the data published by the Center of Excellence for Document Analysis and Recognition at the State University of New York at Buffalo (CEDAR). The data set contains images of handwritten digits with a resolution of 8×8 pixels. There are 11,000 training images and 2,711 images in the test set.

We call the data modified CEDAR data because the data is based on the subsets chosen in [19] and also the preprocessing performed by the authors. We used exactly the same data as Hinton et al. which was also used by Tipping and Bishop (whom we would like to thank for providing the modified data) in [51]. Table 7 gives an overview of the error rates obtained on these data using various methods.

A.5 The ImageCLEF 2005/IRMA Task

The IRMA database contains medical image data from the IRMA project (Image Retrieval in Medical Applications[5]) of the RWTH Aachen University [33]. The database contains 10,000 images that are labeled with a detailed code that specifies body region, image modality, biosystem imaged, and imaging orientation. The data was used as a part of the 2005 ImageCLEF workshop for the evaluation of image retrieval systems.[6] The data set is partitioned into 9,000 training images and 1,000 test images. For the ImageCLEF 2005 task, the data set was subdivided into 57 classes. Although the original images are much larger for the experiments the images were scaled to a common height of 32 pixels to make the computations faster and to allow for an unchanged filter setup. At this scale, fine textures and noise are strongly reduced in the images which allows to use pixel-based gradients.

In Table 8, an overview of the results of the 2005 ImageCLEF "Automatic Annotation Task" is given. For each group, only the best and the worst result among the submissions is included. The complete table is available online.[7] In total, 26 groups registered for participation in the automatic annotation task. From these 26 groups, 12 groups submitted 41 runs, each group had at least two different submissions, the maximum number of submissions per group was 7. A more detailed analysis and a short description of the methods that were used by the groups can be found in [8].

5. http://www.irma-project.org.
6. http://ir.shef.ac.uk/imageclef2005/.
7. http://www-i6.informatik.rwth-aachen.de/~deselaers/imageclef05_aat_results.html.

REFERENCES

[1] E. Alpaydin and C. Kaynak, "Cascading Classifiers," *Kybernetika*, vol. 4, pp. 369-374, 1998.
[2] S. Belongie, J. Malik, and J. Puzicha, "Shape Matching and Object Recognition Using Shape Contexts," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 24, no. 4, pp. 509-522, Apr. 2002.
[3] C.M. Bishop and J.M. Winn, "Non-Linear Bayesian Image Modelling," *Proc. Sixth European Conf. Computer Vision*, pp. 3-17, June 2000.
[4] L. Bottou, C. Cortes, J.S. Denker, H. Drucker, I. Guyon, L. Jackel, Y. Le Cun, U. Müller, E. Säckinger, P. Simard, and V.N. Vapnik, "Comparison of Classifier Methods: A Case Study in Handwritten Digit Recognition," *Proc. Int'l Conf. Pattern Recognition*, pp. 77-82, Oct. 1994.

[5]   D.J. Burr, "Elastic Matching of Line Drawings," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 3, no. 6, pp. 708-713, Nov. 1981.

[6]   D.J. Burr, "A Dynamic Model for Image Registration," *Computer Graphics and Image Processing,* vol. 15, pp. 102-112, Feb. 1981.

[7]   J. Cai and Z.Q. Liu, "Integration of Structural and Statistical Information for Unconstrained Handwritten Numeral Recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 21, no. 3, pp. 263-270, Mar. 1999.

[8]   P. Clough, H. Müller, T. Deselaers, M. Grubinger, T. Lehmann, J. Jensen, and W. Hersh, "The CLEF 2005 Cross-Language Image Retrieval Track," *Proc. CLEF 2005 Workshop Working Notes,* Sept. 2005.

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

[10] D. DeCoste and B. Schölkopf, "Training Invariant Support Vector Machines," *Machine Learning,* vol. 46, nos. 1-3, pp. 161-190, 2002.

[11] J.X. Dong, "Speed and Accuracy: Large-Scale Machine Learning Algorithms and Their Applications," PhD thesis, Dept. of Computer Science, Concordia Univ., Montreal, Canada, 2003.

[12] J.X. Dong, A. Krzyzak, and C.Y. Suen, "Statistical Results of Human Performance on USPS Database," technical report, CENPARMI, Concordia Univ., Montreal, Canada, Oct. 2001.

[13] J.X. Dong, A. Krzyzak, and C.Y. Suen, "A Practical SMO Algorithm," *Proc. Int'l Conf. Pattern Recognition,* vol. 3, Aug. 2002.

[14] H. Drucker, R. Schapire, and P. Simard, "Boosting Performance in Neural Networks," *Int'l J. Pattern Recognition Artificial Intelligence,* vol. 7, no. 4, pp. 705-719, 1993.

[15] C.A. Glasbey and K.V. Mardia, "A Review of Image Warping Methods," *J. Applied Statistics,* vol. 25, no. 2, pp. 155-171, 1998.

[16] B. Haasdonk, A. Halawani, and H. Burkhardt, "Adjustable Invariant Features by Partial Haar-Integration," *Proc. 17th Int'l Conf. Pattern Recognition,* pp. 769-774, Aug. 2004.

[17] B. Haasdonk and D. Keysers, "Tangent Distance Kernels for Support Vector Machines," *Proc. 16th Int'l Conf. Pattern Recognition,* vol. II, pp. 864-868, Sept. 2002.

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

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

[20] G.E. Hinton, C.K.I. Williams, and M. Revow, "Adaptive Elastic Models for Hand-Printed Character Recognition," *Advances in Neural Information Processing Systems 4,* pp. 512-519, 1992.

[21] D. Keysers, J. Dahmen, T. Theiner, and H. Ney, "Experiments with an Extended Tangent Distance," *Proc. 15th Int'l Conf. Pattern Recognition,* vol. 2, pp. 38-42, Sept. 2000.

[22] D. Keysers, T. Deselaers, and H. Ney, "Pixel-to-Pixel Matching for Image Recognition Using Hungarian Graph Matching," *Proc. DAGM Symp. Pattern Recognition,* pp. 154-162, Aug. 2004.

[23] D. Keysers, C. Gollan, and H. Ney, "Classification of Medical Images Using Non-Linear Distortion Models," *Proc. Bildverarbeitung für die Medizin,* pp. 366-370, Mar. 2004.

[24] D. Keysers, C. Gollan, and H. Ney, "Local Context in Non-Linear Deformation Models for Handwritten Character Recognition," *Proc. Int'l Conf. Pattern Recognition,* vol. IV, pp. 511-514, Aug. 2004.

[25] D. Keysers, W. Macherey, H. Ney, and J. Dahmen, "Adaptation in Statistical Pattern Recognition Using Tangent Vectors," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 26, no. 2, pp. 269-274, Feb. 2004.

[26] D. Keysers, R. Paredes, H. Ney, and E. Vidal, "Combination of Tangent Vectors and Local Representations for Handwritten Digit Recognition," *Proc. Int'l Workshop Statistical Pattern Recognition,* pp. 538-547, Aug. 2002.

[27] D. Keysers and W. Unger, "Elastic Image Matching Is NP-Complete," *Pattern Recognition Letters,* vol. 24, nos. 1-3, pp. 445-453, Jan. 2003.

[28] H.J. Kim, D. Kim, and S.Y. Bang, "A Numeral Character Recognition Using the PCA Mixture Model," *Pattern Recognition Letters,* vol. 23, nos. 1-3, pp. 103-111, Jan. 2002.

[29] D.E. Knuth, *The Stanford GraphBase: A Platform for Combinatorial Computing.* Addison-Wesley, 1994.

[30] S. Kuo and O. Agazzi, "Keyword Spotting in Poorly Printed Documents Using Pseudo 2D Hidden Markov Models," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 16, no. 8, pp. 842-848, Aug. 1994.

[31] Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel, "Handwritten Digit Recognition with a Back-Propagation Network," *Advances in Neural Information Processing Systems 2,* pp. 396-404, 1990.

[32] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-Based Learning Applied to Document Recognition," *Proc. IEEE,* vol. 86, no. 11, pp. 2278-2324, Nov. 1998.

[33] T. Lehmann, M. Güld, C. Thies, B. Fischer, K. Spitzer, D. Keysers, H. Ney, M. Kohnen, H. Schubert, and B. Wein, "Content-Based Image Retrieval in Medical Applications," *Methods of Information in Medicine,* vol. 43, no. 4, pp. 354-361, Oct. 2004.

[34] E. Levin and R. Pieraccini, "Dynamic Planar Warping for Optical Character Recognition," *Proc. Int'l Conf. Acoustics, Speech, and Signal Processing,* vol. 3, pp. 149-152, Mar. 1992.

[35] C.L. Liu, K. Nakashima, H. Sako, and H. Fujisawa, "Handwritten Digit Recognition: Benchmarking of State-of-the-Art Techniques," *Pattern Recognition,* vol. 36, no. 10, pp. 2271-2285, Oct. 2003.

[36] D.G. Lowe, "Object Recognition from Local Scale-Invariant Features," *Proc Int'l Conf. Computer Vision,* pp. 1150-1157, Sept. 1999.

[37] R. Marée, P. Geurts, J. Piater, and L. Wehenkel, "A Generic Aproach for Image Classification Based on Decision Tree Ensembles and Local Sub-Windows," *Proc. Asian Conf. Computer Vision,* K.S. Hong and Z. Zhang, eds., vol. 2, pp. 860-865, Jan. 2004.

[38] G. Mayraz and G. Hinton, "Recognizing Handwritten Digits Using Hierarchical Products of Experts," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 24, no. 2, pp. 189-197, Feb. 2002.

[39] C.J. Merz, P.M. Murphy, and D.W. Aha, "UCI Repository of Machine Learning Databases," Univ. of California, Dept. of Information and Computer Science, Irvine, 1997, http://www.ics.uci.edu/~mlearn/MLRepository.html.

[40] J. Milgram, R. Sabourin, and M. Cheriet, "Combining Model-Based and Discriminative Approaches in a Modular Two-Stage Classification System: Application to Isolated Handwritten Digit Recognition," *Electronic Letters on Computer Vision and Image Analysis,* vol. 5, no. 2, pp. 1-15, 2005.

[41] S. Mori, K. Yamamoto, and M. Yasuda, "Research on Machine Recognition of Handprinted Characters," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 6, no. 4, pp. 386-405, July 1984.

[42] F. Perronnin, J.L. Dugelay, and K. Rose, "Iterative Decoding of Two-Dimensional Hidden Markov Models," *Proc. Int'l Conf. Acoustics, Speech, and Signal Processing,* vol. 3, pp. 329-332, Apr. 2003.

[43] M.A. Ronee, S. Uchida, and H. Sakoe, "Handwritten Character Recognition Using Piecewise Linear Two-Dimensional Warping," *Proc. Int'l Conf. Document Analysis and Recognition,* pp. 39-43, Sept. 2001.

[44] B. Schölkopf, *Support Vector Learning.* Oldenbourg Verlag, 1997.

[45] P. Simard, D. Steinkraus, and J.C. Platt, "Best Practices for Convolutional Neural Networks Applied to Visual Document Analysis," *Proc. Seventh Int'l Conf. Document Analysis and Recognition,* pp. 958-962, Aug. 2003.

[46] P. Simard, Y.L. Cun, and J. Denker, "Efficient Pattern Recognition Using a New Transformation Distance," *Advances in Neural Information Processing Systems 5,* pp. 50-58, Apr. 1993.

[47] S.J. Smith, M.O. Bourgoin, K. Sims, and H.L. Voorhees, "Handwritten Character Classification Using Nearest Neighbor in Large Databases," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 16, no. 9, pp. 915-919, Sept. 1994.

[48] L.N. Teow and K.F. Loe, "Handwritten Digit Recognition with a Novel Vision Model that Extracts Linearly Separable Features," *Proc. Conf. Computer Vision and Pattern Recognition,* vol. 2, pp. 76-81, June 2000.

[49] L.N. Teow and K.F. Loe, "Robust Vision-Based Features and Classification Schemes for Off-Line Handwritten Digit Recognition," *Pattern Recognition,* vol. 35, no. 11, pp. 2355-2364, Nov. 2002.

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

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

[52] S. Uchida and H. Sakoe, "A Monotonic and Continuous Two-Dimensional Warping Based on Dynamic Programming," *Proc. Int'l Conf. Pattern Recognition,* vol. 1, pp. 521-524, Aug. 1998.

[53] S. Uchida and H. Sakoe, "Eigen-Deformations for Elastic Matching Based Handwritten Character Recognition," *Pattern Recognition,* vol. 36, no. 9, pp. 2031-2040, Sept. 2003.

[54] S. Uchida and H. Sakoe, "A Survey of Elastic Matching Techniques for Handwritten Character Recognition," *IEICE Trans. Information and Systems,* vol. 88, no. 8, pp. 1781-1790, 2005.

[55] M. Umeda, "Advances in Recognition Methods for Handwritten Kanji Characters," *IEICE Trans. Information and Systems,* vol. 79, no. 5, pp. 401-410, 1996.

[56] P. Viola and W. Wells, "Alignment by Maximization of Mutual Information," *Int'l J. Computer Vision,* vol. 24, no. 2, pp. 137-154, 1997.

[57] L. Wiskott, J. Fellous, N. Kruger, and C. von der Malsburg, "Face Recognition by Elastic Bunch Graph Matching," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 19, no. 7, pp. 775-779, July 1997.

[58] M. Zahedi, D. Keysers, T. Deselaers, and H. Ney, "Combination of Tangent Distance and Image Distortion for Appearance-Based Sign Language Recognition," *Proc. DAGM Symp. Pattern Recognition,* pp. 401-408, Aug. 2005.

Daniel Keysers received the Dipl. degree (with honors) in 2000 and the Dr. degree (summa cum laude) in 2006, both in computer science, from the RWTH Aachen University, Germany. During his PhD studies, he was with the Department of Computer Science of the RWTH and headed the image processing and understanding group at the Human Language Technology and Pattern Recognition Chair. Since 2005, he has been a senior researcher at the German Research Center for Artificial Intelligence (DFKI), Image Understanding and Pattern Recognition Group (IUPR), in Kaiserslautern, Germany. He visited the Instituto Tecnológico de Informática at the Unversidad Politécnica de Valencia, Spain, in 2002 and Microsoft Live Labs in 2006. His research interests include pattern recognition and statistical modeling, especially for computer vision, image object recognition, image retrieval, and document processing.

Thomas Deselaers received the Dipl. degree in computer science (with honors) in 2004 from RWTH Aachen University, Germany. In 2002, he was a visiting student researcher at the Instituto Tecnológico de Informática at the Unversidad Politécnica de Valencia, Spain. Since March 2004, he has been a research assistant and PhD student with the Department of Computer Science of the RWTH Aachen University, where he has been head of the image processing and understanding group at the Human Language Technology and Pattern Recognition Chair since 2005. His research interests are object classification and detection in complex scenes, content-based image retrieval, and pattern recognition. He is a student member of the IEEE.

Christian Gollan studied computer science at the RWTH Aachen University, Germany. He received the Dipl. degree in computer science from the RWTH Aachen in 2004. Since then, he has been a PhD student in the Department of Computer Science of the RWTH, working on automatic speech recognition. His scientific interests cover all aspects of large vocabulary continuous speech recognition, especially unsupervised training.

Hermann Ney received the Dipl. degree in physics from the University of Goettingen, Germany, in 1977 and the Dr.-Ing. degree in electrical engineering from the TU Braunschweig (University of Technology), Germany, in 1982. In 1977, he joined Philips Research Laboratories (Hamburg and Aachen, Germany), where he worked on various aspects of speaker verification, isolated and connected word recognition, and large vocabulary continuous-speech recognition. In 1985, he was appointed head of the Speech and Pattern Recognition Group. In 1988-1989, he was a visiting scientist at AT&T Bell Laboratories, Murray Hill, New Jersey. In July 1993, he joined RWTH Aachen (University of Technology), Germany, as a professor of computer science. His work is concerned with the application of statistical techniques and dynamic programming for decision-making in context. His current interests cover pattern recognition and the processing of spoken and written language, in particular, signal processing, search strategies for speech recognition, language modeling, automatic learning, and language translation. He is a member of the IEEE.

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/publications/dlib.