This paper was published in
Journal of Electronic Imaging 12(1), pp. 59–68 (January 2003)
and is made available as an electronic reprint with permission of SPIE.

Journal of Electronic Imaging 12(1), 59–68 (January 2003).

Statistical framework for model-based image retrieval in medical applications

Daniel Keysers
Jörg Dahmen
Hermann Ney
RWTH Aachen—University of Technology
Lehrstuhl für Informatik VI
Computer Science Department
D-52056 Aachen, Germany
E-mail: keysers@cs.rwth-aachen.de

Berthold B. Wein
RWTH Aachen–University of Technology
Department of Diagnostic Radiology
Medical Faculty
D-52057 Aachen, Germany

Thomas M. Lehmann
RWTH Aachen–University of Technology
Institute of Medical Informatics
Medical Faculty
D-52057 Aachen, Germany

Abstract. *Recently, research in the field of content-based image retrieval has attracted a lot of attention. Nevertheless, most existing methods cannot be easily applied to medical image databases, as global image descriptions based on color, texture, or shape do not supply sufficient semantics for medical applications. The concept for content-based image retrieval in medical applications (IRMA) is therefore based on the separation of the following processing steps: categorization of the entire image; registration with respect to prototypes; extraction and query-dependent selection of local features; hierarchical blob representation including object identification; and finally, image retrieval. Within the first step of processing, images are classified according to image modality, body orientation, anatomic region, and biological system. The statistical classifier for the anatomic region is based on Gaussian kernel densities within a probabilistic framework for multiobject recognition. Special emphasis is placed on invariance, employing a probabilistic model of variability based on tangent distance and an image distortion model. The performance of the classifier is evaluated using a set of 1617 radiographs from daily routine, where the error rate of 8.0% in this six-class problem is an excellent result, taking into account the difficulty of the task. The computed posterior probabilities are furthermore used in the subsequent steps of the retrieval process.* © 2003 SPIE and IS&T. [DOI: 10.1117/1.1525790]

1 Introduction

The importance of digital image retrieval techniques is increasing in the emerging fields of medical imaging and picture archiving and communication systems (PACS). Up to now, textual index entries were necessary to retrieve medical images from a hospital archive, even if the archive was digital imaging and communications in medicine (DICOM) compliant. Currently, much research is done in the field of image retrieval, but the majority of today's content-based image retrieval (CBIR) approaches is intended for browsing large databases of arbitrary content [QBIC (Ref. 1), Photobook,[2] Blobworld[3]], e.g., collected from the World Wide Web.[4] For thorough collections of techniques, we refer to special issues of notable journals, such as *IEEE Transactions on Pattern Analysis and Machine Intelligence* (vol. 18, no. 8, 1996), *IEEE Transactions on Knowledge and Data Engineering* (vol. 10, no. 6, 1998), *Computer Vision and Image Understanding* (vol. 75, no. 1–2, 1999), and *Image and Vision Computing* (vol. 17, no. 7, 1999), or comparative surveys, e.g., Refs. 5 and 6.

Usually, the features used for indexing characterize the entire image rather than image regions or objects, and one of the most effective features of such systems is color.[7] Unfortunately, color-based features are not suitable for the majority of medical images, which are usually gray-scale images. Advanced CBIR approaches are reported (e.g., Ref. 8), but their applicability to medical images remains to be shown. Resulting from the variety of images, common CBIR systems have only a rudimentary understanding of image content, with little or no distinction between important and negligible features or between different anatomical or biological objects in the image. But queries of diagnostic

Paper MIP-06 received May 1, 2001; revised manuscript received Oct. 16, 2002; accepted for publication Jun. 1, 2002.
1017-9909/2003/$15.00

relevance include searching for organs, their relative locations, and other distinct features such as morphological appearances. Therefore, common CBIR systems cannot guarantee a meaningful query completion when used in a medical context. In contrast to this, the image retrieval in medical applications (IRMA) system—a joint project of three RWTH Aachen University of Technology institutes—is being developed for use in daily clinical routine.[9]

This paper presents the general, multistep approach of the IRMA project and reports details of the image classification step within the IRMA system. This first step is of major importance as the retrieval system must know the anatomical region presented in a given image to be able to answer complex medical queries. We present a general probabilistic framework for object recognition and show its effectiveness for radiograph classification, where invariance to small image transformations is incorporated by using invariant distance measures. We also present a thorough quantitative analysis of the performance of the classifier, which is rarely provided in the literature. The analysis is based on 1617 radiographs arbitrarily selected from clinical routine.

2 IRMA System

2.1 *Medical Constraints to Image Retrieval and Related Work*

Since global color, texture, or shape analyses are insufficient to characterize medical images, retrieval results are rather poor when common CBIR systems are applied to medical images.[2,10,11] In recent reports, some approaches for content-based retrieval designed to support specific medical tasks have been published. Korn *et al.* describe a system for fast and effective retrieval of tumor shapes in mammogram x rays,[12] where the morphological features are defined on binary images. To transfer this promising approach to other tasks, redesign of the structuring elements is required. One emphasis of their work is fast searching in the underlying database, which we do not consider explicitly in this paper. Other feature-extraction methods or similarity models are known, where especially invariant features are used for image retrieval.[8,13] For instance, shape histograms are used in Ref. 14 in combination with a quadratic-form distance function comparable to the Mahalanobis distance with a structured covariance matrix. The automatic search and selection engine with retrieval tools (ASSERT) system operates on high resolution computed tomographies of the lung.[15] A physician delineates the region bearing a pathology and marks a set of anatomical landmarks when the image is entered into the database. Hence, ASSERT has high data entry costs, which prohibit its application in clinical routine. Chu *et al.* present a knowledge-based retrieval system with spatial and temporal constructs.[16] Brain lesions are automatically extracted within three-dimensional (3-D) data sets from computed tomography and magnetic resonance imaging. Their representation model consists of an additional knowledge-based layer within the semantic model. This layer provides a mechanism for accessing and processing spatial, evolutionary, and temporal queries. However, those concepts for medical image retrieval are task-specific and not directly transferable to other medical applications.

Tagare *et al.* point out some of the unique challenges retrieval engines are confronted with when dealing with medical image collections.[17] Medical knowledge arises from anatomic and physiologic information, requiring regional features to support diagnostic queries. However, interpretation of medical images depends on both image and query context. Since the context of queries is unknown when images are entered into the database, the database scheme must be generic and flexible. Particularly, the number and kind of features are subject to continuous evolution. Furthermore, categorization and registration of medical images are required to support diagnostic queries on a high level of image interpretation.

2.2 *The Multistep Approach*

In contrast to common approaches to image retrieval, the IRMA concept is based on a separation of the following seven steps to enable complex image content understanding[9] (Fig. 1):

1. image categorization (based on global features)
2. image registration (in geometry and contrast)
3. feature extraction (using local features)
4. feature selection (category and query dependent)
5. indexing (multiscale blob representation)
6. identification (incorporating prior knowledge)
7. retrieval (on abstract blob level)

2.2.1 *Categorization*

Various imaging techniques require adapted image processing methods. For example, ultrasonic images of vessels must be processed in a different manner than skeletal radiographs. Thus, if a radiologist is searching the database for all radiographs showing a pulmonal tumor, the IRMA system processes only radiographs that have a sufficiently high posterior probability for the class "chest." Therefore, the categorization step not only reduces the computational complexity of an IRMA query, it will also reduce the false-alarm rate of the system, improving its precision. Based on global features, the IRMA approach distinguishes four major categories:

1. image modality (physical)
2. body orientation (technical)
3. anatomic region (anatomical)
4. biological system (functional)

These categories build subclasses resulting in hierarchically structured categories.[9] Thus far, only the first level of the anatomical hierarchy is used for classification experiments.

Modern modalities enable submission of textual information about the examination. However, the medical staff often does not enter appropriate or sufficient data into the systems, as a recent study showed quantitatively [only one out of four examined modalities included the correct digital imaging and communication in medicine (DICOM) header information and even in this case the information was in-

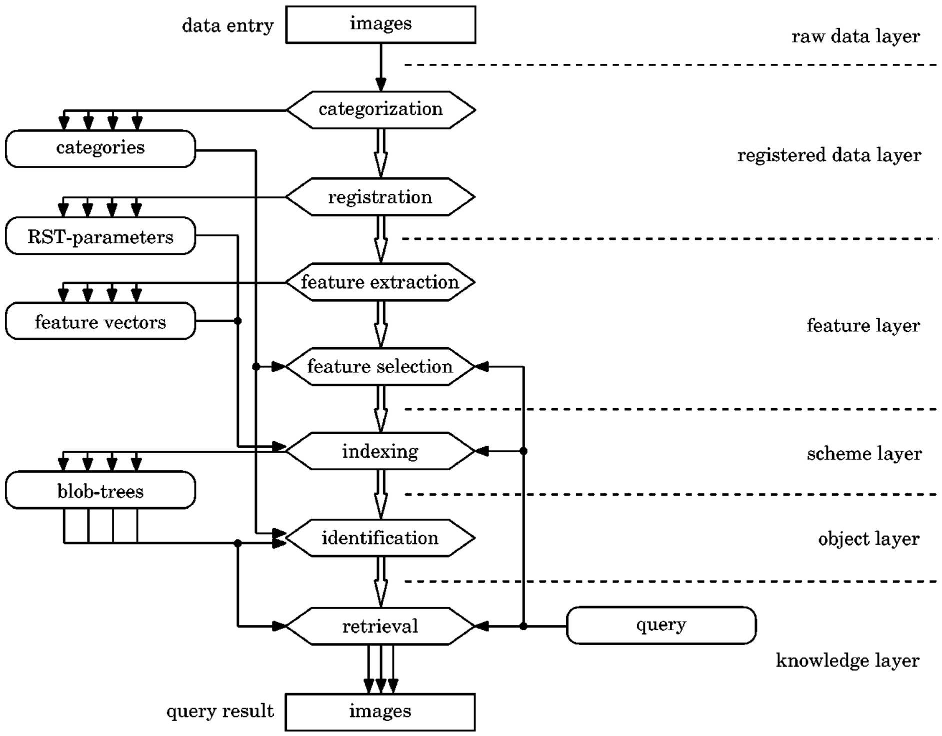

Fig. 1 IRMA architecture.

correct in 15.5% of the examined cases].[18] In a well-managed DICOM-compliant PACS linked to a hospital information system, text-based retrieval will give excellent results. But in many cases automatic indexing by image content is still a necessary component to provide sufficient information, where these two methods should not be viewed as mutually exclusive but as synergistic tools.

2.2.2 *Registration*

Diagnostic inferences derived from images are deduced from an incomplete but continuously evolving model of normality.[17] In the IRMA system, this model is represented by prototype images, which are defined for each category by an expert based on prior medical knowledge or by statistical analysis. The prototypes are used for determination of parameters for rotation, scaling, and translation as well as contrast adjustment. However, the images are not transformed at this stage of processing.

2.2.3 *Feature extraction*

The processing of semantic queries drawn from medical routine requires local features, e.g., the local gradient, which are connected to pixels. Like the global features for categorization, the number of local feature images is extensible. Category-specific local features include segmentation by active contours or active shapes,[19] which enable the use of prior shape knowledge from the categories.

2.2.4 *Feature selection*

It is important that feature extraction and selection are separate steps in the IRMA concept. This enables the latter task to be retrieval-dependent. Prior knowledge about the image category as well as medical knowledge incorporated into the query is used to select a precomputed set of adequate feature images. For example, the retrieval of radiographs with respect to bone fractures or bone tumors is done using an edge-based or texture-based feature set, respectively.

2.2.5 *Indexing*

For query processing, the amount of information processed in the previous steps must be drastically reduced. Based on feature sets, the image is segmented hierarchically into relevant regions, which are described by invariant moments resulting in structures called blobs.[3] Thereafter, the blob representation of the image is adjusted with respect to the parameters determined in the registration step.

2.2.6 *Identification*

The hierarchical blob tree has been registered with respect to a certain category, where categories are represented by prototypes. Since the local features reflect characteristic and discriminant properties of tissue, certain blobs correspond to well defined morphological structures in the image. Vice versa, prior medical knowledge on the content and structure of category prototypes can be used to build a prototype blob structure with characteristic properties for blob identification and labels for semantic queries. While blob identification probably will not be successful for all blob entities, blob trees are an applicable data structure to incorporate medical knowledge into the IRMA system.

2.2.7 *Retrieval*

Retrieval is performed by searches in the hierarchical blob structures. In IRMA, a query is built from the following components:

1. a list of possible categories of the recall images
2. a query by example blob-structure on the optimal scale to process the query
3. the set of local features that best describe significant properties for the query

However, not all of these components must be set for all queries. The assignment of categories to the images significantly increases the performance of retrieval as only the blob structures for images of the possible categories have to be compared. However, data structures and distance measures that have been described for image retrieval[20] require refining so that only selected blobs and features are relevant in the query.

2.3 *Interdependence of the Steps*

Care must be taken with (possibly false) local decisions because of the interdependence of the steps. To improve the possibility of obtaining the overall best possible query result, it is necessary to work with a variety of different hypotheses throughout these steps. This holistic approach has shown superior performance over local decisions in many applications, such as speech recognition.[21] An example of modeling of vague knowledge is the computation of posterior probabilities during the categorization step, which avoids the hard choice of a possibly false image category. Thus, it is not necessary to correct a false decision later, but instead the entire process works on multiple hypotheses at the same time. It is helpful to consider classification methods for CBIR because classification and image retrieval aim at similar objectives, which has been pointed out before (e.g., in Ref. 11).

3 Image Database

The image database used in the experiments consists of medical radiographs taken from daily routine (with the patient information erased), which are secondary digital, i.e., they have been scanned from conventional film-based radiographs (Fig. 2). The corpus consists of 110 abdomen, 706 limbs, 103 breast, 110 skull, 410 chest, and 178 spine radiographs (200×200 up to 2500×2500 pixels, 8 bits), summing to a total of 1617 images.[22] The distribution of the images reflects the clinical routine at the Department of Diagnostic Radiology, RWTH Aachen University Hospital, Germany. The quality of radiographs varies considerably and there is a great within-category variability (as caused by different doses of x rays, varying orientations, images with and without pathologies or contrast agents, changing scribor position, etc.). Furthermore, there is a strong visual similarity between many images of the classes abdomen and spine (Fig. 2). Therefore, the classification problem can be considered being hard.

Furthermore, a smaller set of 332 images exists that is used to test the generalization abilities of the classifier. To speed up the classification process, the original images are scaled down to a common height of 32 pixels (keeping the original aspect ratio). In previous experiments, it was shown that this downscaling does not affect classification performance significantly.[22] Since there are only 1617 images available, we apply a leaving-one-out approach for cross validation. That is, to classify an image we use the remaining 1616 images as references and report results averaged over the entire 1617 experiments.

Fig. 2 Example radiographs taken from the IRMA database, scaled to common height: left to right, top row, abdomen, limbs, breast, middle row, skull, chest, and spine; bottom row, examples of variation within one class (limbs).

4 Feature Extraction

We make use of appearance-based pattern recognition, i.e., each pixel of an image is interpreted as a feature. Thus, all the information contained in an image is used for classification. As an additional feature throughout the experiments we use the aspect ratio of the images. Although invariances play an important role for classifying radiographs, we do not extract invariant features. Instead, we incorporate these invariances in the classification algorithm itself. This is done using distance measures that are—for instance, invariant to transforms such as image rotation, axis deformations, scaling or varying image brightness. In the following, we denote images of dimension $I\times J$ by $\mathbf{x}=\{x_{ij}\in\mathbb{R}:i=1,\ldots,I,j=1,\ldots,J\}$ with x_{ij} being the gray level at pixel (i,j).

5 Statistical Framework and Classification

The first step of the IRMA concept is a classification task. To classify an observation $\mathbf{x}\in\mathbb{R}^{I\times J}$ we use the Bayesian decision rule[23]

$$\mathbf{x}\mapsto r(\mathbf{x})=\operatorname*{argmax}_k\{p(k)p(\mathbf{x}|k)\}, \tag{1}$$

where $p(k)$ is the prior probability of class k and $p(\mathbf{x}|k)$ is the class-conditional probability for the observation $\mathbf{x}$ given class k (Ref. 24). The decision rule [Eq. (1)] is known to be optimal with respect to the number of classification errors (assuming equal cost for each error), if the true probability density functions are known.[23] For multiple-object recognition, we extend the elementary decision rule into the following directions:

1. We assume that the scene $\{x_{ij}\}$ contains an unknown number M of objects belonging to the classes $k_1,\ldots,k_M=:k_1^M$. Reference models $p(\mathbf{x}|\mu_k)$ exist for each of the classes $k=1,\ldots,K$, and μ_0 represents the background.
2. We make decisions about object boundaries, i.e., the original scene is implicitly partitioned into $M+1$ regions I_0^M, where $I_m \subset \{(i,j):i=1,\ldots,I,j=1,\ldots,J\}$ is assumed to contain the m'th object, and I_0 represents the background.
3. The reference models may be subject to certain transforms (rotation, scale, translation, etc.). That is, given transformation parameters ϑ_1^M, the m'th reference is mapped to $\mu_{k_m} \rightarrow \mu_{k_m}(\vartheta_m)$.

The idea is now to consider all unknown parameters, i.e., M,k_1^M,ϑ_1^M, and (implicitly) I_0^M and to search the hypothesis which best explains the given scene. This must be done considering the interdependence between the image partitioning, where partitioning is only part of the classification process (holistic concept). Note that this means that any pixel in the scene must be assigned either to an object or to the background class. The resulting decision rule is

$$r(\{x_{ij}\}) = \underset{M,k_1^M,\vartheta_1^M}{\operatorname{argmax}} \left\{ p(k_1^M) \prod_{m=0}^{M} p[\mathbf{x}_{I_m}|\mu_{k_m}(\vartheta_m)] \right\}, \tag{2}$$

where $\{x_{ij}\}$ denotes the scene to be classified, and $\mathbf{x}_{I_m}$ is the feature vector extracted from region I_m. Note that if not searching for all of the parameters, a summation over the disregarded parameters should be performed. As the true density functions are not known in practical situations, we must choose models for these functions and estimate their parameters from the training data. Invariance aspects are directly incorporated into these models using a probabilistic model of variability.[24] In Eq. (2), $p(k_1^M)$ is a prior over the combination of objects in the scene, which may depend on the transformation parameters and the combination of objects (e.g., a skull located close to a spine is more likely than close to a foot).

The consideration of all the components of the presented decision rule [Eq. (2)] is a long-term goal. We started with the consideration of the interdependence between segmentation and recognition. For classification of medical images, the "scene" equals the radiograph to be classified and we assume $M=1$. Thus, Eq. (2) reduces to

$$r(\{x_{ij}\}) = \underset{k,\vartheta}{\operatorname{argmax}} \{p(k)p(\mathbf{x}_{I_0}|\mu_0)p[\mathbf{x}_{I_1}|\mu_k(\vartheta)]\}. \tag{3}$$

Furthermore, the only transformation considered for the reference images in the experiments is horizontal shift (vertical shifts do not occur as all images are scaled to the same height and objects are assumed to be centered). A simple background model is used for p_0, assuming a constant background of grayvalue zero and large variance. Based on the different sizes of observation and reference images, a penalty term is introduced preferring images of roughly the same size. To model the references $p[\mathbf{x}_{I_1}|\mu_k(\vartheta)]$, kernel densities are used:

$$p[\mathbf{x}|\mu_k(\vartheta)] = \frac{1}{N_k} \sum_{n=1}^{N_k} \frac{1}{A_{k,\gamma}} \exp\left\{ -\frac{d[\mathbf{x},\mu_{kn}(\vartheta)]}{\sigma_k^2 \cdot \gamma} \right\}, \tag{4}$$

where N_k is the number of reference images of class k, μ_{kn} is the n'th reference pattern of class k, σ_k denotes the class specific standard deviation and $d[\mathbf{x},\mu(\mu_{kn},\vartheta)]$ one of the (squared) distance measures introduced later (which are not necessarily metrics). To compensate for the fact that variances are usually underestimated if only few training samples are available, we multiply the estimated variances with a factor $\gamma > 1$. Strictly speaking, the normalization factor $A_{k,\gamma}$ depends on the class k, however, the dependency is weak and therefore neglected in the experiments. The prior probabilities are modeled using the relative frequencies $p(k) = N_k/N$ with the total number of reference images N. Similar approaches to model the distribution of feature vectors are common practice, also in the processing of medical images. For example, a combinatorial search to find the set of most discriminative features for a kernel density classifier for brain images is applied in Ref. 11.

Resulting from the exponential decay with increasing distance in Eq. (4), only the closest reference patterns μ_{kn} contribute substantially to the sum. Using more than the 10 closest matches with respect to d does therefore usually not change classification results significantly. To determine these closest matches, it is generally sufficient to compute d for the (e.g., 100) closest references in terms of Euclidean distance, which can be efficiently determined.

6 Invariance

The kernel density classifier is based on invariant distance measures. The baseline for the experimental results is here the distance corresponding to a Gaussian density, i.e., the Mahalanobis distance.[23] Because of the high dimensionality of the extracted feature vectors (e.g., for a region of size 32 $\times$ 32 we have 1024 features) in comparison to the number of samples per class we used a multiple $\sigma_k^2 \cdot \gamma \cdot I$ of the identity matrix as class specific covariance matrix in the experiments. Note that the tangent distance can be interpreted in a probabilistic framework,[25] implying that the term "Gaussian kernel densities" is still applicable for tangent distance when the tangent vectors are based on the references. Strictly speaking, when using the image distortion model, the density functions are not Gaussian any more.

6.1 *Tangent Distance*

In 1993, Simard *et al.* proposed an invariant distance measure called tangent distance (TD), which proved to be es-

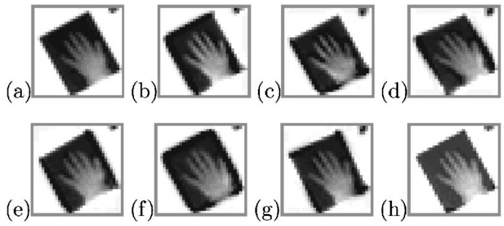

Fig. 3 Example images generated using linear approximations of affine transforms and image brightness: (a) original image, (b) left shift, (c) down shift, (d) hyperbolic diagonal deformation, (e) hyperbolic axis deformation, (f) scaling, (g) right rotation, and (h) increased brightness.

pecially effective in the domain of optical character recognition.[26] The authors observed that reasonably small transforms of certain image objects do not affect the class membership. Simple distance measures such as the Euclidean distance do not account for this. Instead, they are very sensitive to affine transformations such as scaling, translation, rotation or axis deformation. When an image $\mathbf{x} \in \mathbb{R}^{I \times J}$ is transformed (e.g., scaled and rotated) by a transformation $t(\mathbf{x},\alpha)$, which depends on L parameters $\alpha \in \mathbb{R}^L$ (e.g., the scaling factor and rotation angle), the set of all transformed patterns,

$$M_{\mathbf{x}} = \{t(\mathbf{x},\alpha) : \alpha \in \mathbb{R}^L\} \subset \mathbb{R}^{I \times J}, \tag{5}$$

is a manifold of at most dimension L in pattern space. The distance between two patterns can now be defined as the minimum distance between their respective manifolds, being truly invariant with respect to the L regarded transforms. However, computation of this distance is a hard nonlinear optimization problem and the manifolds concerned generally do not have an analytic expression. Therefore, small transformations of the pattern $\mathbf{x}$ are approximated by a tangent subspace $\widehat{M_{\mathbf{x}}}$ to the manifold $M_{\mathbf{x}}$ at the point $\mathbf{x}$. This subspace is obtained by adding to $\mathbf{x}$ a linear combination of the vectors $\mathbf{v}_l, l=1,\ldots,L$ that are the partial derivatives of $t(\mathbf{x},\alpha)$ with respect to α_l and span the tangent subspace. We obtain a first-order approximation of $M_{\mathbf{x}}$:

$$\widehat{M_{\mathbf{x}}} = \left\{ \mathbf{x} + \sum_{l=1}^{L} \alpha_l \cdot \mathbf{v}_l : \alpha \in \mathbb{R}^L \right\} \subset \mathbb{R}^{I \times J}. \tag{6}$$

The one-sided tangent distance $d_{\mathrm{TD}}(\mathbf{x},\mu)$ is then defined as

$$d_{\mathrm{TD}}(\mathbf{x},\mu) = \min_\alpha \left\{ \left\| \mathbf{x} + \sum_{l=1}^{L} \alpha_l \cdot \mathbf{v}_l - \mu \right\|^2 \right\}. \tag{7}$$

The tangent vectors $\mathbf{v}_l$ can be computed using finite differences between the original image $\mathbf{x}$ (or μ) and a small transformation of that image.[26] In the experiments, we computed the six tangent vectors for affine transforms (two translations, rotation, scaling and two axis deformations) as proposed by Simard *et al.*, but replaced the tangent vector for "line thickness" (which is important for optical character recognition) by a "brightness" tangent vector (Fig. 3). All elements of this vector were set to a constant value to

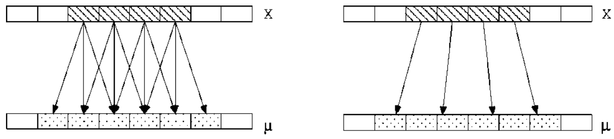

Fig. 4 One-dimensional comparison of IDM (left) and TD (right) (Ref. 27).

model the image brightness, which is determined by the x-ray dose and other factors. A two-sided TD can also be defined, where both manifolds are approximated and the distance is minimized over possible combinations of the respective parameters. Furthermore, the approximation of the manifolds can be improved by an iterative procedure based on Newton's method, which is computationally more expensive.[26,27]

6.2 *Image Distortion Model*

Although TD alone is already a very effective means to compensate for small global transformations of an image, it is highly sensitive to local image transformations. These are, e.g., caused by noise, pathologies, varying collimator fields, or changing scribor positions in a radiograph. We therefore propose the following image distortion model (IDM). When calculating the distance between two images $\mathbf{x}$ and μ, local deformations are allowed, i.e., the "best-fitting" pixel in the reference image within a certain neighborhood R_{ij} is regarded instead of computing the squared error between x_{ij} and μ_{ij}. Figure 4 shows a 1-D example for the IDM where individual pixel displacements are independent compared to the TD, where displacements are coupled forming an affine transform (here scaling). The resulting distance is

$$d_{\mathrm{IDM}}(\mathbf{x},\mu) = \sum_{i,j} \min_{(i',j') \in R_{ij}} \{\|x_{ij} - \mu_{i'j'}\|^2 + C_{iji'j'}\}. \tag{8}$$

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

Note that the TD and the proposed distortion model can be easily combined to a so-called distorted TD. In that case, the tangent distance is used to register the (sub)images and the distortion distance is then computed between the registered images.[24,28]

Table 1 Error rates in percentages for the IRMA database (statistical approach with kernel densities using leaving-one-out).

	This Work	
	Thresholding	
Distance measure	No	Yes
Mahalanobis distance	14.0	11.2
TD	13.3	11.1
IDM	12.1	9.0
Distorted TD	10.4	8.0
Results for Comparison		
Squared images, 1-*NN*		18.1
Squared images, kernel densities		16.4
+ aspect ratio		14.9
Cooccurrence matrices		29.0

6.3 *Relating TD and IDM*

It is interesting to see that the positive effects of the TD and IDM are additive in the case of radiograph classification. When trying to relate these two approaches, it becomes clear that one can be expressed in terms of the other.[27] Generalizing the IDM yields

$$d_{C,\mathcal{F}}(\mathbf{x},\mu)=\min_{f\in\mathcal{F}}\left[C(f)+\sum_{i,j}\|x_{ij}-\mu_{f(i,j)}\|^2\right], \tag{9}$$

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

$$\mathcal{F}_{\mathrm{IDM}}=\{f:f(i,j)\in R_{ij}\},\quad C_{\mathrm{IDM}}(f)=\sum_{i,j}C_{ijf(i,j)}, \tag{10}$$

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

$$\mathcal{F}_{\mathrm{TD}}=\{f:f\text{ affine}\},\quad C_{\mathrm{TD}}(f)=0. \tag{11}$$

This general expression is an intuitive representation of a distance being invariant to arbitrary functions f of some class $\mathcal{F}$. Computing Eq. (9) may be computationally expensive with some classes and cost functions (e.g., for the warping model presented in Ref. 29 no polynomial minimization algorithm is known), but the TD and IDM are two examples with known effective solutions. [In the case of TD this is true at least for a reasonable approximation since strictly speaking, Eq. (11) models the true manifold distance.]

7 Experimental Results

The experimental results are summarized in Table 1. As the images are scaled to the same height, the possible regions that are hypothesized in Eq. (2) are restricted to rectangular areas of the same height, which enables efficient maximization. The baseline results were obtained using the Mahalanobis distance, resulting in an error rate of 14.0%. Using the presented image distortion model with a region size of 3×3 pixels, the error rate was reduced to 12.1%. Although the distortion model is straightforward, it effectively compensates for local variations in radiographs. Using only the TD, the error rate was 13.3%. This gain was not as large as for the distortion model, but is still remarkable. In another experiment, it was investigated whether the improvements of the TD and the IDM are additive. This sounds reasonable, as the TD compensates for global image transformations, whereas the IDM deals with local perturbations. Indeed, using the distorted TD, the error rate was further reduced to 10.4%. As a few large differences in pixel values can mislead classifiers based on squared error distances (e.g., Ref. 30), a local threshold was introduced to limit the maximum contribution of a single pixel difference to the distance between two images. This is especially justified here, because the images may be subject to artifacts that do not affect class membership, such as noise or changing scribor position in radiographs. Applying this thresholding approach, the error rate was reduced to 8.0%.

To make sure that no overfitting occurred in the experiments, the 332 previously unseen radiographs were used as test images and the 1617 images of the IRMA database as references. Using the optimal parameters for the database determined by the leaving-one-out strategy, the algorithm misclassified 30 of the new radiographs (9.0%), which means that the classifier proposed here generalizes very well.

Regarding the computational complexity, the distance functions require a different amount of computation in comparison to the Mahalanobis distance (which is a multiple of the Euclidean distance here). For the one-sided tangent distance the tangents for the prototypes can be calculated before classification. Thus, only the projection into the subspace must be calculated in the recognition step, which increases the computational cost approximately by a factor of $(L+1)$. If the tangent vectors are calculated on the basis of the observations, the cost for the computation and orthogonalization of the tangent subspace must be added. In the IDM, the number of comparisons of pixel values is increased roughly proportional to the size of the regarded region R_{ij}.

In the course of this work, various other experiments were carried out, some of which are worth mentioning. For example, several tangents based on other transformations (e.g., projective) were tested in experiments, but no improvement over the combination of affine and brightness transforms was obtained. Furthermore, experiments concerning the creation of virtual data (a method that was very successful for the task of optical character recognition[27]) did not yield improvements on this particular dataset.

The use of cooccurrence matrices[31] is often considered to be helpful for content-based medical image retrieval. However, our experiments on radiograph classification do not support this thesis. In two experiments, we used global cooccurrence matrices for feature analysis within a synergetic classifier[32] and within a kernel density based classifier. In both cases, we were not able to obtain classification error rates below 29%. Apparently, cooccurrence matrices do not provide discriminative features for radiograph classification. Nevertheless they might still be useful for the subsequent IRMA steps, e.g., for tumor localization within a (previously categorized) radiograph. In this case, cooccurrence matrices would be computed from small parts of the image, not from the complete image.[33]

8 Discussion

8.1 *IRMA Concept*

Figure 1 summarizes the multistep approach of the IRMA system. The seven processing steps are sequentially combined, where the interdependence of local decisions must be regarded as mentioned above. These processing steps correspond to five semantic layers for knowledge representation. Like other systems, the unprocessed images form the first layer, which is called the raw data layer. Categorization (which is presented in detail here) and registration within each category are the first level where medical knowledge is incorporated into the IRMA system. Hence, both steps result in the second semantic layer, which is called the registered data layer. While other medical CBIR systems are restricted to a certain modality or diagnostic procedure,[12,15,16] the registered data layer in IRMA enables queries across all kind of medical images regardless of modality, orientation, body region, or biological system. Further semantic layers are the feature layer, the scheme layer, the object layer, and the knowledge layer.[9] Since retrieval is performed on the knowledge layer, all other layers are processed at data entry time. Hence, they are not critical for the system performance. In general, the IRMA concept is related to the Blobworld project.[3] However, there are several important extensions of the Blobworld concept especially designed for medical purposes,[9] which allow prior knowledge on both image and query content to be used for content-based image indexing.

8.2 *Database*

The size of the database used with currently less than two thousand images, may seem small in comparison with large-scale databases with millions of entries that must to be handled in real-world PACS. Nevertheless, to the best of our knowledge, the IRMA database is the first database of this size containing medical images from daily routine that are labeled by an expert, enabling a detailed evaluation of the classification performance. On larger databases it will be necessary to reduce the computational load using known algorithms for the access of large databases.[7,12,14]

8.3 *Statistical Framework*

Once the maximizing arguments in the decision rule [Eq. (2)] have been determined, it is straightforward to construct a parse tree as a description of the image from the implicit segmentation information, which is done using a neural net approach in Ref. 34. Special attention to the subject of occlusion is paid in Ref. 35, where mainly object contours are considered for recognition, not the objects themselves. Some considerations with respect to a statistical model for multiple images can also be found in Ref. 36. Here, the author concentrates on determining the unknown 3-D transformation parameters in the recognition process as well as improving feature extraction. Localization was improved by explicit modeling of the background, although a global optimization was not performed in the experiments. This approach has recently been extended.[37] Since the optimization determines the computational complexity of the method, iterative and locally optimal algorithms such as the expectation maximization algorithm can also be considered.[38] Note that the framework presented here does not impose any restriction on the type of feature extraction used. This ensures extensibility to multimodal datasets, where aligned images can be treated as one image with an extended feature vector per pixel. Furthermore, the extension to multidimensional images is straightforward, but the search space that must considered grows rapidly with the number of dimensions. However, the resulting difficulties can be handled by using efficient search strategies.

8.4 *Transformation Models*

Conceptionally, both TD and IDM can be extended to multimodal and multidimensional datasets. Multimodality (again assuming a reasonable registration) can be treated using extended feature vectors. Whereas more than the considered two dimensions do not require changes in the tangent model and for the IDM only the considered regions must be adjusted. Nevertheless, the applicability of the approach to such datasets remains to be evaluated experimentally.

Concerning the generalized transformation model [Eq. ((9)], it is interesting to investigate which other cases (besides TD and IDM) may be useful for invariant pattern recognition, and whether one can learn the functions efficiently from training examples. For instance, the IDM can be extended naturally to introduce a dependency between the displacements of pixels in a neighborhood, such that displacements in the same direction are "cheaper" than those in opposite directions. This leads to more complex minimization problems, where one example with interesting properties but high computational complexity is 2D warping[29] and its extensions.[39]

8.5 *Results*

The result for the TD using a local threshold is only slightly better than that of Mahalanobis distance (11.1 versus 11.2%). A possible explanation for this behavior is that using the thresholding approach may mimic the behavior of the TD in this particular application, because the subspace projection minimizes the sum of squared pixel differences. Note also that in previous experiments, all IRMA images were scaled down to a common size of 32×32 pixels prior to classification (more information on that approach is given in Refs. 22 and 24). In these experiments, the TD significantly outperformed the Mahalanobis distance (with and without the thresholding approach). Thus, it seems possible that the main effect of the TD is the compensation of image shifts (which is now inherent in the classification approach by optimizing over all possible image positions). Interestingly, the background model with independent pixel assignments used in Ref. 37 also results in local thresholding and can be interpreted as its probabilistic justification.

9 Conclusions

A statistical framework for model-based IRMA was presented. Based on a classification of the image content, the IRMA concept provides a high amount of content understanding and enables highly differentiated queries on an abstract information level. Furthermore, the concept fulfills the demands for medical image retrieval systems postulated by Tagare *et al.* and therefore, it promises satisfactory query completion.[17] Note, however, that the concept remains to be tested and so far only parts of it have been evaluated in detail.

We presented an approach to statistical classification of radiographs, which is applied in the first step of the IRMA system. We introduced a probabilistic framework for (multi-) object recognition and proved its effectiveness by applying it to radiograph classification (being a single-object recognition task), obtaining an excellent result. Invariance was incorporated into the appearance-based approach by using invariant distance measures. We proposed an effective IDM to compensate for local transformations and motivated a combination with the TD, which is well-suited for global transformations.

The classifier was evaluated by applying it to the IRMA database of radiographs, consisting of 1617 images of six major body regions taken from daily routine. A thorough quantitative analysis of classification like the one presented here is rarely found for medical image retrieval. The best classification error rate of 8.0% was achieved by using the distorted TD in a kernel density classifier within the statistical framework. This is a relative improvement of 42% with respect to the baseline statistical system with 14.0% error rate and a relative improvement of 56% with respect to the error rate of 18.1% obtained by a nearest-neighbor classifier using the Euclidean distance.

Future work on the categorization step will include improved background models and extensions of the algorithm with respect to scale and other transformations. Furthermore, extensions of the generalized distortion model with respect to regularization properties will be considered. In addition, more global features will be considered in combination with a finer level of the hierarchical structure of categories.

The very satisfying results obtained on the hard image classification task can be regarded a solid basis for the further development of the IRMA system.

Acknowledgment

This work was partially funded by the Deutsche Forschungsgemeinschaft (DFG Grant No. Le 1108/4-1).

References

1. W. Niblack, R. Barber, W. Equitz, M. Flickner, E. Glasman, D. Pektovic, P. Yanker, C. Faloutsos, and G. Taubin, "The QBIC project: querying images by content using color, texture, and shape," in *Storage and Retrieval for Image and Video Databases*, *Proc. SPIE* **1908**, 173–187 (1993).
2. A. Pentland, R. Picard, and S. Sclaroff, "Photobook: content-based manipulation of image databases," in *Storage and Retrieval for Image and Video Databases*, *Proc. SPIE* **2185**, 34–47 (1994).
3. C. Carson, M. Thomas, S. Belongie, J. Hellerstein, and J. Malik, "Blobworld: a system for region-based image indexing and retrieval," in *Proc. 3rd Int. Conf. Visual Information Systems*, Vol. 1614 of LNCS, pp. 509–516, Springer, Amsterdam, The Netherlands (June 1999).
4. J. R. Smith and S.-F. Chang, "Tools and techniques for color image retrieval," in *Storage and Retrieval for Image and Video Databases*, *Proc. SPIE* **2670**, 426–437 (1996).
5. M. de Marsico, L. Cinque, and S. Levialdi, "Indexing pictorial documents by their content: a survey of current techniques," *Image Vis. Comput.* **15**(2), 119–141 (1997).
6. A. W. M. Smeulders, M. Worring, S. Santint, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Trans. Pattern Anal. Mach. Intell.* **22**, 1349–1380 (2000).
7. C. Faloutsos, R. Barber, M. Flickner, J. Hafner, W. Niblack, D. Petkovic, and W. Equitz, "Efficient and effective querying by image content," *J. Intell. Inf. Syst.* **3**, 231–262 (1994).
8. C. Schmid and R. Mohr, "Local grayvalue invariants for image retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.* **19**(5), 530–535 (1997).
9. T. Lehmann, B. Wein, J. Dahmen, J. Bredno, F. Vogelsang, and M. Kohnen, "Content-based image retrieval in medical applications: a novel multi-step approach," *Proc. SPIE* **3972**(32), 312–320 (2000).
10. M. Nappi, G. Polese, and G. Tortora, "FIRST: fractal indexing and retrieval system for image databases," *Image Vis. Comput.* **16**(14), 1019–1031 (1998).
11. Y. Liu and F. Dellaert, "A classification based similarity metric for 3D image retrieval," in *Proc. Int. Conf. Computer Vision and Pattern Recognition*, pp. 800–805, IEEE, Santa Barbara, CA (June 1998).
12. F. Korn, N. Sidiropoulos, C. Faloutsos, E. Siegel, and Z. Protopapas, "Fast and effective retrieval of medical tumor shapes," *IEEE Trans. Knowl. Data Eng.* **10**(6), 889–904 (1998).
13. H. Burkhardt and S. Siggelkow, "Invariant features in pattern recognition—fundamentals and applications," in *Nonlinear Model-Based Image/Video Processing and Analysis,* C. Kotropoulos and I. Pitas, Eds., pp. 269–307, Wiley, New York (2001).
14. M. Ankerst, G. Kastenmüller, H.-P. Kriegel, and T. Seidl, "3D shape histograms for similarity search and classification in spatial databases," in *Proc. 6th Int. Symp. on Spatial Databases*, Vol. 1651 of *LNCS*, pp. 207–226, Springer, Hong Kong, China (July 1999).
15. C. Shyu, C. Brodley, A. Kak, A. Kosaka, A. Aisen, and L. Broderick, "ASSERT: a physician-in-the-loop content-based image retrieval system for HRCT image databases," *Comput. Vis. Image Underst.* **75**(1–2), 111–132 (1999).
16. W. W. Chu, C. C. Hsu, A. F. Cárdenas, and R. K. Taira, "Knowledge-based image retrieval with spatial and temporal constructs," *IEEE Trans. Knowl. Data Eng.* **10**(6), 872–888 (1998).
17. H. Tagare, C. Jaffe, and J. Duncan, "Medical image databases: a content-based retrieval approach," *J. Am. Med. Inform Assoc.* **4**(3), 184–198 (1997).
18. M. Kohnen, H. Schubert, B. B. Wein, R. W. Günther, J. Bredno, T. M. Lehmann, and J. Dahmen, "Qualität von DICOM-Informationen in Bilddaten aus der klinischen Routine," in *Bildverarbeitung für die Medizin 2001*, pp. 419–423, Springer, Lübeck, Germany (Mar. 2001).
19. F. Weiler and F. Vogelsang, "Model-based segmentation of hand radiographs," *Proc. SPIE* **3338**(1), 673–681 (1998).
20. J. Goldstein, R. Ramakrishnan, U. Shaft, and J. B. Yu, "Using constraints to query R⋆-trees," Technical report 1301, Dept. of Computer Science, University of Wisconsin, Madison (Feb. 1996).
21. H. Ney and S. Ortmanns, "Progress in dynamic programming search for LVCSR," *Proc. IEEE* **88**(8), 1224–1240 (2000).
22. J. Dahmen, T. Theiner, D. Keysers, H. Ney, T. Lehmann, and B. Wein, "Classification of radiographs in the 'image retrieval in medical applications' system (IRMA)," in *Proc. 6th Int. RIAO Conf. on Content-Based Multimedia Information Access,* pp. 551–566, Paris, France (Apr. 2000).
23. R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd. ed., Wiley, New York (2001).
24. J. Dahmen, D. Keysers, H. Ney, and M. O. Güld, "Statistical image object recognition using mixture densities," *J. Math. Imaging Vision* **14**(3), 285–296 (2001).
25. D. Keysers, W. Macherey, J. Dahmen, and H. Ney, "Learning of variability for invariant statistical pattern recognition," in *ECML 2001, 12th Eur. Conf. on Machine Learning*, Vol. 2167 of LNCS, pp. 263–275, Springer, Freiburg (Sep. 2001).
26. P. Simard, Y. Le Cun, and J. Denker, "Efficient pattern recognition using a new transformation distance," in *Advances in Neural Information Processing Systems 5*, S. Hanson, J. Cowan, and C. Giles, Eds., pp. 50–58, Morgan Kaufmann, San Mateo, CA (1993).
27. D. Keysers, J. Dahmen, T. Theiner, and H. Ney, "Experiments with an extended tangent distance," in *Proc. 15th Int. Conf. on Pattern Recognition*, Vol. 2, pp. 38–42, IEEE, Barcelona, Spain (Sep. 2000).
28. J. Dahmen, D. Keysers, M. Motter, H. Ney, T. Lehmann, and B. Wein, "An automatic approach to invariant radiograph classification," in *Bildverarbeitung für die Medizin 2001*, pp. 337–341, Springer, Lübeck, Germany (Mar. 2001).
29. S. Uchida and H. Sakoe, "A monotonic and continuous two-dimensional warping based on dynamic programming," in *Proc. 14th Int. Conf. on Pattern Recognition*, Vol. 1, pp. 521–524, IEEE, Brisbane, Australia (Aug. 1998).
30. N. Vasconcelos and A. Lippman, "Multiresolution tangent distance for affine-invariant classification," in *Advances in Neural Inf. Proc. Systems*, M. I. Jordan, M. J. Kearns, and S. A. Solla, Eds., Vol. 10, pp. 843–849, MIT Press, Cambridge, MA (1998).
31. R. Haralick, K. Shanmugam, and I. Deinstein, "Textural features for image classification," *IEEE Trans. Syst. Man Cybern.* **3**(6), 610–621 (1973).

32. H. Haken, *Synergetic Computers and Cognition*, Springer Verlag, New York (1991).
33. F. Weiler, F. Vogelsang, M. Kilbinger, B. Wein, and R. W. Günther, ''Automatic recognition of image contents using textural information and a synergetic clasifier,'' *Proc. SPIE* **3034**, 985–989 (1997).
34. G. Hinton, Z. Ghahramani, and Y. Teh, ''Learning to parse images,'' in *Advances in Neural Information Processing Systems 12*, S. Solla, T. Leen, and K. Müller, Eds., pp. 463–469, MIT Press, Cambridge, MA (2000).
35. K. Mardia, W. Qian, D. Shah, and K. de Souza, ''Deformable template recognition of multiple occluded objects,'' *IEEE Trans. Pattern Anal. Mach. Intell.* **19**(9), 1035–1042 (1997).
36. J. Pösl and H. Niemann, ''Wavelet features for statistical object localization without segmentation,'' in *Proc. Int. Conf. on Image Processing*, pp. 170–173, IEEE, Santa Barbara, CA (1997).
37. M. Reinhold, D. Paulus, and H. Niemann, ''Appearance-based statistical object recognition by heterogeneous background and occlusions,'' in *Pattern Recognition, 23rd DAGM Symp.*, Vol. 2191 of *LNCS*, pp. 254–261, Springer, Munich, Germany (2001).
38. S. Akaho, ''The EM algorithm for multiple object recognition,'' in *Proc. Int. Conf. on Neural Networks ICNN'95*, pp. 2426–2431, IEEE, Perth, Australia (Nov. 1995).
39. S. Uchida and H. Sakoe, ''Piecewise linear two-dimensional warping,'' in *Proc. 15th Int. Conf. on Pattern Recognition,* Vol. 3, pp. 538–541, IEEE, Barcelona, Spain (Sep. 2000).

Daniel Keysers studied computer science at the Aachen University of Technology (RWTH), Germany, and at the Universidad Complutense de Madrid, Spain. He received his Dipl degree in computer science (with honors) from the RWTH Aachen in 2000 and he has since been a PhD student with the Department of Computer Science, working on statistical methods and models for image object recognition, including aspects of invariance.

Jörg Dahmen received his Dipl degree in computer science from the Aachen University of Technology (RWTH), Germany, in June 1997. In August 1997, he joined the Department of Computer Science, RWTH, as a PhD student, where his research interests covered various aspects of image processing and statistical pattern recognition, especially object recognition in images. He received his PhD degree in computer science from the RWTH Aachen in 2002.

Hermann Ney received his Dipl degree in physics from the University of Göttingen, Germany, in 1977 and his 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 and 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 for computer science. His work concerns 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.

Berthold B. Wein studied human medical sciences at the Aachen University of Technology (RWTH), Germany. He is currently the chief consultant with the Department of Radiology, the University Medical Center in Aachen, and a professor of diagnostic radiology. His main scientific interests are image processing, computer vision and reasoning, databases, and organizational systems (picture archiving and communication systems, radiology information systems, and health information systems).

Thomas M. Lehmann received his MS degree in electrical engineering and his PhD in computer science from the Aachen University of Technology (RWTH), Germany, in 1992 and 1998, respectively. He heads the Department of Medical Image Processing at the Institute of Medical Informatics, RWTH Aachen. In 1993 he received an award from the German Association for Pattern Recognition (DAGM-Preis '93). In 1998 he received the Borcher's Medal from the RWTH Aachen. He is member of IEEE, SPIE, and IADMFR, and chairs the IEEE Joint Chapter Engineering in Medicine and Biology (IEEE German Section). Since 1999, he has served on the international editorial board of *Dentomaxillofacial Radiology*. His research interests are medical image processing applied to quantitative measurements for diagnoses and content-based image retrieval from large medical databases. He has authored several papers and a textbook on medical image processing (Springer-Verlag, Berlin) and edited a handbook on medical informatics (Hanser-Verlag, Munich).