ELSEVIER

Computerized Medical Imaging and Graphics 29 (2005) 143–155

Computerized Medical Imaging and Graphics

www.elsevier.com/locate/compmedimag

Automatic categorization of medical images for content-based retrieval and data mining

Thomas M. Lehmann[a,*], Mark O. Güld[a], Thomas Deselaers[b], Daniel Keysers[b], Henning Schubert[c], Klaus Spitzer[a], Hermann Ney[b], Berthold B. Wein[c]

[a]*Department of Medical Informatics, Medical Faculty, Aachen University of Technology (RWTH), Pauwelsstr. 30, Aachen D-52057, Germany*
[b]*Chair of Computer Science VI, Department of Computer Science, Aachen University of Technology (RWTH), Aachen, Germany*
[c]*Department of Diagnostic Radiology, Medical Faculty, Aachen University of Technology (RWTH), Aachen, Germany*

Received 26 March 2004; revised 29 August 2004; accepted 30 September 2004

Abstract

Categorization of medical images means selecting the appropriate class for a given image out of a set of pre-defined categories. This is an important step for data mining and content-based image retrieval (CBIR). So far, published approaches are capable to distinguish up to 10 categories. In this paper, we evaluate automatic categorization into more than 80 categories describing the imaging modality and direction as well as the body part and biological system examined. Based on 6231 reference images from hospital routine, 85.5% correctness is obtained combining global texture features with scaled images. With a frequency of 97.7%, the correct class is within the best ten matches, which is sufficient for medical CBIR applications.

Keywords: Content-based image retrieval (CBIR); Data mining; Medical imaging; Pattern recognition; Feature extraction; Image categorization; Texture analysis; Classifier combination

1. Introduction

The increasing amount of digitally produced images requires new methods to archive and access this data. Conventional databases allow for textual searches on meta data only. Often, the database scheme only holds references to the image data, which are stored as individual files on the file system. Especially images may contain semantic information that cannot be conveyed by a textual description. Thus, a growing interest in image data mining and content-based image retrieval (CBIR) can be observed [1].

While data mining denotes the analysis of (often large) observational data sets in order to find unsuspected relationships and to summarize the data in ways that are better understandable to human observers [2], CBIR aims at searching image databases for specific images that are similar to a given query image. Here, the search is based on the appearance of the images instead of (or in addition to) using a textual description. Usually, a sample image is presented to the system, which answers this query by returning all similar matches. This concept is referred to as the query by example (QBE) paradigm. It was introduced by Niblack et al. when presenting the query by image content (QBIC) system in the early 1990s [3,4]. Consequently, CBIR and the QBE paradigm do not directly aim at summarizing data. Rather, they are concerned with the understandable presentation of relevant extracts of a large set of data to a user. However, the methods studied in CBIR are mostly relevant to data mining in image databases, as many data mining processes rely on notions of similarity or distance between data items, which are also used and investigated in CBIR research.

There are several areas of application for CBIR systems. For instance, biomedical informatics compiles large image databases. In particular, medical imagery is increasingly

* Corresponding author. Tel.: +49 241 80 88793; fax: +49 241 80 3388793.
E-mail address: lehmann@computer.org (T.M. Lehmann).
URL: www.http://irma-project.org/lehmann.

0895-6111/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compmedimag.2004.09.010

acquired, transferred, and stored digitally. In large hospitals, several tera bytes of data need to be managed each year [5]. However, picture archiving and communication systems (PACS) still provide access to the image data by alphanumerical description and textual meta information. This also holds for digital systems compliant with the Digital Imaging and Communications in Medicine (DICOM) protocol. Therefore, integrating CBIR into medicine is expected to significantly improve the quality of patient care [6].

Müller et al. have recently reviewed the increasing research on content-based retrieval approaches to medical applications [5]. The majority of research is focussed on a particular image content, modality, body region, or pathology. Therefore, categorization of medical images is important for medical CBIR systems that are not restricted to a specific context. Especially in applications of digital radiology such as computer-aided diagnosis or case-based reasoning, the image category is of major importance for subsequent processing steps because it allows context-specific selection of appropriate filters or algorithmic parameters.

Categorization of medical images means image classification into a predefined order scheme. For instance, the Systemized Nomenclature of Medicine (SNOMED, http://www.snomed.org), the Medical Subject Heading (MeSH, http://nlm.nih.gov/mesh), as well as the Unified Medical Language Systems (UMLS, http://nlm.nih.gov/research/umls) provide order codes for the body region examined, the imaging modality used, and contrast agents applied for the examination. But so far, the categorization is usually done manually by the physician or radiologist during the routine documentation. The DICOM header also provides tags to decode the body part examined and the patient position, which are usually set by the digital modality according to the imaging protocol used to capture the pixel data. However, this information cannot always be considered reliable [7]. Therefore, automatic and reliable categorization of medical images is an important field of research.

This paper is organized as follows. In Section 2, we briefly review CBIR systems provided for medical applications. In particular, the semantics of features that are used in such systems is analyzed. In Section 3, we focus on medical image categorization using global texture features. The exhaustive experiments that are presented in Section 4 are based on a large set of more than 6000 images arbitrarily selected from clinical routine and representing as many as 80 categories. Based on the promising results (Section 5), a perspective of medical CBIR is given in Section 6.

2. Content-based image retrieval in medicine

Fig. 1 shows the general system architecture for content-based retrieval. At data entry time, numerical features are computed from each image stored within the database. Using the QBE approach, the same features are extracted from the query image and compared to the features stored within the database. The images that correspond to the most similar features are then retrieved from the database and presented to the user to answer his query.

Each module in Fig. 1 also exemplifies where the approaches of medical CBIR systems can be distinguished. Most systems apply several restrictions to the input images [5]. As mentioned before, in the majority of published papers, the systems are unable to cope with images of more than one modality or body region, since they implicitly or explicitly use particular properties of the data during image processing. For instance, the WebMIRS system proposed by Long et al. combines text-based retrieval on the meta data stored in a database with content-based retrieval on x-rays of the human cervical and lumbar spines [9]. The images are roughly segmented, the anatomical segments are labelled by structural names, and finally these regions are classified according to pathology or high-level semantic features of interest. A high amount of a priori knowledge about the structure of the images is used for these preprocessing steps. Specific systems for computed tomographies (CT) of the head, functional positron emission tomographies (PET), mammographies, or photographies obtained in dermatology are overviewed in [5].

Also, the Medical CBIR systems differ in query formulation. Although the QBE paradigm is most prominent, manually drawn sketches or shapes are also used to input a query [10]. From the technical point of view, a large

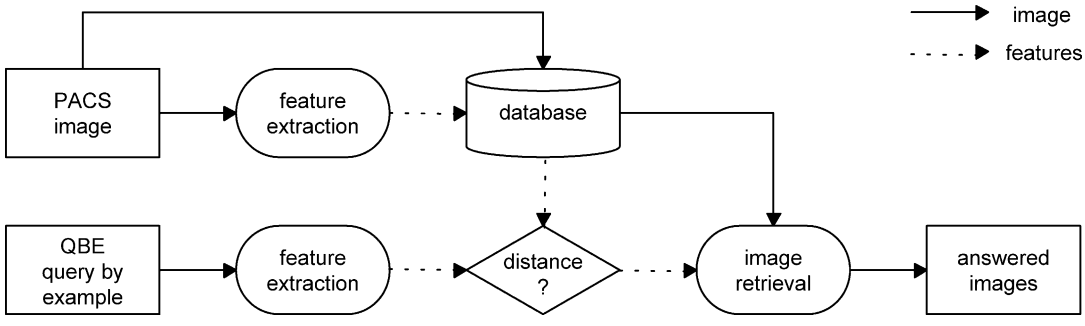

Fig. 1. Architecture of CBIR systems.

variety is found with respect to database organization (e.g. distributed or central storage, indexing and fast access methods), design of the retrieval engine (e.g. static or learning from the user's interaction, providing query refinement and relevance feedback), and concepts of user interfacing (e.g. proprietary or web-based front-ends, complexity and usability).

However, the most important differences result from the feature extraction methods and the distance measures that are used to compute the most similar responses with respect to a given query (Fig. 1). In the majority of systems, the images are described by global features. Here, a single numerical value or a small set of numbers, which are then combined into a single feature vector, are used to represent the entire image. In general CBIR approaches, these numbers are computed from global color histograms. In medical applications, color is rather seldom present and less discriminative as compared to general image collections, i.e. as obtained from the internet. Furthermore, most information in medical images is local [6]. This opens the so called semantic gap, which is defined as 'the lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data has for a user in a given situation [1].'

In order to close the semantic gap, the image retrieval in medical applications (IRMA) approach defines successive semantic levels of information abstraction, which are sequentially computed [8,11]. In the first stage of feature extraction, global texture descriptions are used only to determine the imaging modality and view direction as well as the body region and biological system imaged. With respect to the image category, appropriate local features are used to partition the image into visually significant regions. Since the level of detail may vary with respect to the query, this segmentation is performed hierarchically by iterative region merging. In the finest level of details, each pixel is represented by its own feature vector while on the top level, the entire image is merged into the root node, which is connected to a single feature vector. As a result, the image is represented by a graph in a tree topology and image similarity is computed by means of graph or subgraph matching. Based on different sets of features, a couple of graphs is stored with each image and individually selected according to the actual context of the user's query. For instance, queries concerning bone fractures or bone tumors within the same skeletal radiograph are answered based on a graph that was computed from edge or texture features, respectively.

According to Müller et al., only five other research projects currently aim at creating a content-based image retrieval system for general medical applications, namely I^2C, COBRA, KMeD, MedGIFT, and ImageEngine [5]. The ImageEngine [12] is primarily a text-based information retrieval system but first experiments produced interesting results in combination with some basic components of computer vision. Both, I^2C [13] and COBRA [14] characterize the images based on global features only. In I^2C, these features are calculated from the complete image and in COBRA, they are extracted from an automatically segmented image region. The MedGIFT system is based on the freely available GNU Image Finding Tool (GIFT) [15]. A hierarchical approach is applied that is based on effective concepts used in textual information retrieval. In particular, a very high dimensional feature space (dimensionality approx. 85,000) of different low-level features is used to compute inverted files that allow efficient access to the data. Closer related to the IRMA approach, KMeD [16] introduces four semantic layers to model local features and their spatial relationship.

Regardless of the number of semantical levels defined by each system, all of them apply global features on a rather low level of semantics, i.e. features that directly describe the appearance of the images compressing the information that is based on the pixel values. Although several works discuss semantic image retrieval [17–20], only some systems attempt small steps towards this goal. For example, these first steps consist of connecting low-level features with textual high-level features [20]. Other systems try to find objects in the images and a mapping of the objects in one image onto the objects in another image [18,19]. This means that a semantically valid segmentation is needed, which is yet an unsolved problem in image processing, as image segmentation is very closely connected to image understanding. Consequently, manual annotations are required in several system approaches to medical CBIR [6].

In summary, the current state of the art of medical CBIR systems is dominated by global low-level features applied for different tasks, which, however, can all be expressed in a short and concise manner by the term image categorization.

3. Categorization of medical images

In general, automatic categorization as a mapping of images into pre-defined classes involves three basic principles [21]:

(i) representation, i.e. the extraction of appropriate features to describe the image content,
(ii) adaptation, i.e. the selection of the best feature subset regarding discriminative information, and
(iii) generalization, i.e. the training and evaluation of a classifier.

So far, automatic categorization of medical images is restricted to a small number of categories. For instance, several algorithms have been proposed for orientation detection of chest radiographs, where lateral and frontal orientation are distinguished by means of digital image processing [22,23]. For this two-class experiment, the error rates are below 1% [24]. In a recent investigation, Pinhas and Greenspan report error rates below 1% for automatic

categorization of 851 medical images into eight classes [25]. In a previous investigation of the IRMA group, six classes are defined according to the body part examined from 1617 images and an error rate of 8% is reported [26].

However, such a low number of classes is not suitable for applications in evidence-based medicine or case-based reasoning. Here, the image category must be determined in much more detail. In the following sections, it is therefore analyzed whether global features can still be used to distinguish medical images into a large number of predefined categories that represent semantics rather than dense clusters in feature space.

4. Experiments on automatic categorization

4.1. The image corpus

A detailed classification scheme has been developed to encode medical images according to their content [27] The four axes of the IRMA code assess the imaging technique and modality (T-axis, four levels of detail), the relative direction of the imaging device and the patient (D-axis, three levels of detail), the anatomic body part that is examined (A-axis, three levels of detail), and the biological system being under investigation (B-axis, three levels of detail). Thus, each image encoding has the form TTTT-DDD-AAA-BBB, with presently 797 unique entities available on the four axes.

Currently, about 10,000 images have been taken randomly from clinical routine at the Aachen University Hospital, Aachen, Germany, and manually IRMA-coded resulting in more than 400 used codes. In contrast to other coding schemes, the IRMA code is mono-hierarchical, i.e. without cycles, which allows to uniquely merge sub-groups. At the date of our experiments, the corpus contained 6335 images. Mostly, secondary digitized images from plain radiography (5839 images) but also images from other modalities, e.g. computed tomography and ultrasound imaging were collected. All images have been categorized by an experienced radiologist according to the IRMA code [27]. In total, 351 different codes were assigned and several codes were used for one or two images, only. Since it is almost impossible to effectively categorize images from categories with very few members available, we take advantage of the IRMA-code hierarchy and pursue 2,1,2, and 1 levels of detail on the T-,D-,A-, and B-axis, respectively. This yields 135 unique IRMA codes matching the scheme TT**-D**-AA*-B**. Additionally, a threshold is applied for the minimum number of images in each category and all images from categories below the threshold are disregarded. This results in 6231 images from 81 categories using a minimum of five images per category (Table 1).

4.2. Image features

As previously mentioned, global features describing color and shape, which are commonly applied in CBIR-systems, are mostly inapplicable in the medical domain. According to previous investigations, we applied texture measures and resized representations of the images as global feature vectors.

4.2.1. Texture measures

Considering texture, a wide range of features has been proposed in the literature. To make the texture properties comparable, all images were scaled into an identical size of 256×256 pixels, ignoring the initial aspect ratio. Based on several experiments, those features being most suitable to distinguish medical images have been chosen:

1. Based on the fundamental work of Haralick and coworkers [28], Tamura et al. suggested coarseness, contrast, and directionality to describe an image's texture properties [29]. These features are computed on a per-pixel basis. Therefore, we collect the values into a three-dimensional histogram ($6\times8\times8=384$ bins) and use the Jensen-Shannon-divergence to measure the similarity between two histograms [30].
2. Castelli et al. used various texture features to describe image properties [31]. These encompass the global fractal dimension (computed using reticular cell counting), the coarseness, the gray-scale histogram entropy, some spatial gray-level statistics, and the circular Moran autocorrelation function. In all, 43 values are extracted and combined into a feature vector.
3. Motivated from fast indexing of JPEG-compressed images, Ngo et al. used the variance of the first nine alternation current (AC) coefficients obtained by the discrete cosine transform (DCT) over all 8×8 pixel blocks of an image [32]. Applied to medical images, results were improved when the direct current (DC) and some more of the AC coefficients are also considered. In this study, the first 21 DCT coefficients are used.
4. In 2001, Zhou and Huang proposed an algorithm to capture properties of edges within an image [33]. A water-filling process is applied to the binarized gradient image. Canny's edge detector is used to determine the gradient. The three parameters, i.e. the deviation of the Gaussian kernel used to smooth the image as well as the lower and the upper threshold for the edge tracing algorithm, were empirically optimized. According to the authors' initial suggestion, we use the filling time, fork count, and loop count, where both counts are computed for a global and a per-edge-segment maximum.

4.2.2. Scaled image representations

In previous work, down-scaled images $r(x,y)$ and $s(x,y)$ have been used successfully as feature vectors. Here, r and s denote the reference and the search image, respectively.

To obtain vectors of identical size $h \times h$, $h \in \{8,16,24,32\}$, the images are again scaled ignoring their original aspect ratio. For this type of feature, several similarity measures can be applied.

1. Resulting from its mathematical simplicity, the Euclidean distance measure (EDM)

$$D_{\mathrm{EDM}}(r,s) = \sqrt{\sum_{x=1}^{h}\sum_{y=1}^{h}(r(x,y)-s(x,y))^2} \tag{1}$$

determines the pixel-wise quadratic distance between both images. However, EDM results in large distances for similar radiographs acquired from the same region of the same patient within the same orientation but with different radiation.

2. Adopted from signal processing, the normalized cross-covariance function (CCF)

$$D_{\mathrm{CCF}}(r,s) = \max_{|m|,|n|\le d}\left\{\frac{\sum_{x=1}^{h}\sum_{y=1}^{h}(r(x-m,y-n)-\bar{r})(s(x,y)-\bar{s})}{\sqrt{\left(\sum_{x=1}^{h}\sum_{y=1}^{h}(r(x-m,y-n)-\bar{r})^2\right)\cdot\left(\sum_{x=1}^{h}\sum_{y=1}^{h}(s(x,y)-\bar{s})^2\right)}}\right\} \tag{2}$$

returns the maximum correlation over a selected warp range, i.e. the two-dimensional translations over d pixels are performed explicitly. In Eq. (2), $\bar{r}$ and $\bar{s}$ denote the pixel-wise mean gray value of r and s, respectively. For our experiments, we used $d = \lfloor h/8 \rfloor$, where $\lfloor \cdot \rfloor$ denotes the truncation of a real into the next lower integer value. Note that CCF normalizes the image brightness, which is another common cause of variability found in medical images.

3. The tangent distance was introduced by SIMARD et al. [34]. It models the manifold for each vector in the feature space generated by small transformations by using a linear approximation, the so-called tangent subspace. The projection of a sample onto a reference's subspace can then be computed efficiently. Let $t(r,\alpha)$ denote a transformation of an image r, which depends on L parameters $\alpha \in \lambda^L$ (e.g. the scaling factor and the rotation angle). Then, the subspace is obtained by a linear combination of the vectors v_l, $l = 1,\ldots,L$ that are the partial derivatives of $t(r,\alpha)$ with respect to α_l, which is added to the resized image $r(x,y)$. The tangent distance D_{TAN} is then defined as the minimum distance between the tangent subspaces of reference and observation (two-sided tangent distance). In the experiments, only one of the two subspaces was considered (one-sided tangent distance) [24]:

$$D_{\mathrm{TAN}}(r,s) = \min_{\alpha}\left\{\sqrt{\sum_{x=1}^{h}\sum_{y=1}^{h}\left(\left(r(x,y)+\sum_{l=1}^{L}\alpha_l v_l(x,y)\right)-s(x,y)\right)^2}\right\} \tag{3}$$

Instead of an approximation for line thickness, as it has been initially proposed by Simard et al. for optical character recognition, a tangent modelling global brightness is integrated. Since there is a trade-off between important image details, which are lost when the representation is too small, and robustness to noise or non-perfect alignments, which is lost once the representation gets too detailed, we chose sizes of $h \in \{16,24,32\}$.

4. Alternatively, the image distortion model (IDM) allows local displacements for each pair of corresponding pixels compared within the distance measure [35]. This is especially useful for medical images due to individual anatomical properties in each image. The policy is to match each pixel of the sample image to one in the reference image. This ensures that all sample pixels are explained by the reference. To prevent a completely unordered vector field of pixel mappings between two images, it is useful to include the local context into the search process for a correspondence hypothesis. Denoting the coordinate offsets by x'' and y'', while x' and y' denote the offsets within the search window for a corresponding pixel, the distance is computed by

$$D_{\mathrm{IDM}}(r,s) = \sum_{x=1}^{X}\sum_{y=1}^{Y}\min_{|x'|,|y'|\le W_1}\left\{\sum_{|x''|,|y''|\le W_2}||r(x+x'+x'', y+y'+y'') - s(x+x'', y+y'')||_2\right\} \tag{4}$$

where X, $Y \le h$ denote the size of the scaled images keeping their original aspect ratio. The results are improved if the horizontal and vertical image gradients as computed by a Sobel filter are used instead of the intensity values. For our experiment, $W_1 = 2$ (5×5 pixel-sized search window for corresponding pixels), $W_2 = 1$ (3×3 pixels of local context), and $h = 32$ are used.

4.3. Automatic classifiers

A k-nearest-neighbor classifier (k-NN) is used, which embeds the distance measures for the features described above. The classifier opts for the category which gets the most votes over the k references that are closest to the sample vector according to the distance measured. In our experiments, k is chosen from $\{1,5\}$. This is a simple yet effective method, which is also useful to present classification results interactively.

According to Jain, classifier combination can be grouped into three main categories [21]:

(i) parallel,
(ii) serial (like a sieve), and
(iii) hierarchical (comparable to a tree).

Since it is an easy way to post-process existing results obtained from the single classifiers, the classifiers are combined parallel. Another reason is that we examine dynamic category partitioning of the image corpus and do not focus on the optimization of one static category set at present. For parallel combination, the classifier results are first transformed to a common scale. Then, a weighted summation of the results is performed to compute the combined classifier vote. For a first experiment, a smaller subset of the image corpus is used to optimize the weighing coefficients, which are then applied to combine the results for the full image corpus.

4.4. Evaluation

Based on the image corpus, exhaustive experiments were carried out using the leaving-one-out scheme for evaluation [36]. Each time, one image is used as the test image and the remaining images as references. Then, the mean categorization rate over all permutations is computed. The hierarchical organization of the code allows to investigate classification results at a certain level of detail (given enough images per category for meaningful experiments).

Table 1
The 81 image categories in use for evaluation

Category index	IRMA code	Images		Category index	IRMA code	Images	
		Absolute	Relative (%)			Absolute	Relative (%)
1	11**-1**-50*-0**	1278	20.51	42	11**-2**-91*-7**	29	0.47
2	11**-2**-50*-0**	611	9.81	43	14**-3**-73*-0**	28	0.45
3	11**-1**-41*-7**	448	7.19	44	14**-3**-21*-7**	28	0.45
4	11**-2**-23*-7**	179	2.87	45	12**-1**-73*-4**	28	0.45
5	11**-1**-20*-7**	174	2.79	46	11**-4**-23*-7**	28	0.45
6	11**-2**-33*-7**	165	2.65	47	11**-2**-44*-7**	28	0.45
7	11**-2**-31*-7**	157	2.52	48	11**-1**-44*-7**	28	0.45
8	11**-4**-21*-7**	155	2.49	49	11**-2**-21*-7**	27	0.43
9	11**-1**-31*-7**	152	2.44	50	11**-1**-42*-7**	25	0.40
10	11**-1**-33*-7**	139	2.23	51	11**-2**-42*-7**	24	0.39
11	11**-1**-80*-7**	124	1.99	52	11**-1**-93*-2**	24	0.39
12	11**-1**-70*-4**	116	1.86	53	31**-2**-94*-7**	23	0.37
13	11**-2**-94*-7**	105	1.69	54	14**-3**-71*-0**	22	0.35
14	11**-1**-94*-7**	101	1.62	55	11**-1**-43*-7**	21	0.34
15	11**-1**-46*-7**	99	1.59	56	11**-2**-93*-7**	20	0.32
16	11**-2**-32*-7**	89	1.43	57	14**-3**-20*-7**	18	0.29
17	11**-4**-62*-6**	87	1.40	58	12**-1**-71*-4**	18	0.29
18	11**-4**-61*-6**	86	1.38	59	31**-2**-30*-7**	16	0.26
19	11**-1**-91*-7**	86	1.38	60	13**-1**-50*-0**	16	0.26
20	11**-4**-41*-7**	85	1.36	61	14**-3**-73*-4**	15	0.24
21	11**-3**-62*-6**	82	1.32	62	12**-1**-52*-3**	15	0.24
22	11**-1**-32*-7**	82	1.32	63	11**-2**-22*-7**	15	0.24
23	11**-3**-61*-6**	80	1.28	64	11**-1**-93*-7**	15	0.24
24	11**-1**-96*-7**	75	1.20	65	11**-1**-71*-4**	15	0.24
25	11**-1**-45*-7**	65	1.04	66	11**-1**-70*-5**	14	0.22
26	13**-1**-80*-2**	64	1.03	67	31**-3**-33*-7**	11	0.18
27	11**-2**-92*-7**	64	1.03	68	31**-3**-32*-7**	11	0.18
28	11**-1**-92*-7**	64	1.03	69	11**-1**-95*-2**	10	0.16
29	11**-2**-41*-7**	63	1.01	70	11**-1**-72*-4**	10	0.16
30	11**-3**-94*-7**	60	0.96	71	12**-1**-72*-4**	9	0.14
31	11**-1**-51*-7**	54	0.87	72	12**-2**-73*-4**	8	0.13
32	14**-3**-20*-1**	52	0.83	73	12**-1**-53*-3**	8	0.13
33	11**-1**-95*-7**	45	0.72	74	11**-1**-73*-4**	8	0.13
34	11**-2**-96*-7**	44	0.71	75	14**-4**-71*-2**	7	0.11
35	11**-1**-21*-7**	43	0.69	76	14**-3**-50*-0**	7	0.11
36	11**-2**-95*-7**	36	0.58	77	11**-1**-80*-2**	7	0.11
37	11**-2**-46*-7**	36	0.58	78	13**-1**-72*-2**	6	0.10
38	14**-3**-72*-0**	34	0.55	79	13**-1**-71*-2**	6	0.10
39	11**-2**-43*-7**	34	0.55	80	12**-1**-70*-4**	5	0.08
40	11**-4**-31*-7**	31	0.50	81	11**-2**-45*-7**	5	0.08
41	11**-5**-91*-7**	29	0.47		****-***-***-***	6231	100.00

The IRMA code is explained in Table 2.

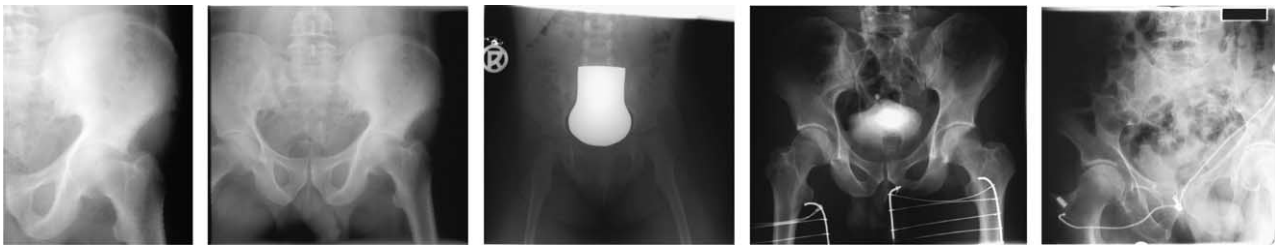

Fig. 2. Intra-class variability. All radiographs are coded identically (IRMA 1121-120-800-700).

Since the IRMA concept proposes to pursue the most likely categories for each unknown image for further content abstraction [8,11], it was also investigated whether the correct category occurs among the first k neighbors. This estimates how many hypotheses must be kept for subsequent processing steps.

5. Results

The feature describing properties of the edge structure performs worst in all experiments and does not exceed 22% recognition rate (Table 1). The texture features proposed by Castelli and the features based on Ngo's approach perform on a similar level of about 40%. Note however, that the DCT-based feature vector contains only half the number of components. For these features, a best recognition rate of 43.9% resulted. The histograms based on Tamura's texture features yielded the best results among the features proposed for general-purpose image retrieval (66% correctness). In nearly all cases, 5-NN improves the recognition rate for this type of feature.

In general, the scaled representations perform better than all texture features examined. Even for Euclidian distance on 8×8 pixel representations, which is the most primitive approach, EDM and CCF yield more that 70%. On very small images, CCF performs worse than EDM. IDM for representations scaled to a fixed height of 32 pixels yields the best results of 82.3%. Contrary to the texture features, the best results among the scaled representations are obtained using 1-NN.

CCF and IDM model spatial variability within a local neighborhood while the texture features capture rather global image properties. Therefore, a combination of classifiers based on the IDM (best among scaled representations) and the texture features according to Tamura (best among global texture features) is evaluated (Table 3). An improvement to over 85% recognition rate is obtained for both, 1-NN and 5-NN classifiers. However, for the task of retrieving the correct class within the ten best matches, the best rate of 97.72% is obtained using re-scaled images of 24×24 pixels with CCF.

Additional experiments for radiographs only (5776 images from 57 categories) and a minimum number of 10 class members (6155 images from 70 categories for all

Table 2
The IRMA codes used in Table 1

Technique	
11**	Plain radiography
12**	Fluoroscopy
13**	Angiography
14**	Computed tomography
31**	Magnetic resonance imaging
Direction	
1**	Coronal
2**	Sagittal
3**	Axial
4**	Other
Anatomy	
20*	Cranium, unspecified
21*	Facial cranium
22*	Cranial base
23*	Neuro cranium
30*	Spine, unspecified
31*	Cervical spine
32*	Thoracic spine
33*	Lumbar spine
41*	Hand
42*	Radio carpal joint
43*	Forearm
44*	Elbow
45*	Upper arm
46*	Shoulder
50*	Chest, unspecified
51*	Chest, bones
52*	Lung
53*	Hilum
61*	Right breast
62*	Left breast
70*	Abdomen, unspecified
71*	Upper abdomen
72*	Middle abdomen
73*	Lower abdomen
80*	Pelvis, unspecified
91*	Foot
92*	Ankle joint
93*	Lower leg
94*	Knee
95*	Upper leg
96*	Hip
Biosystem	
0**	Unspecified
1**	Cerebrospinal system
2**	Cardiovascular system
3**	Respiratory system
4**	Gastrointestinal system
5**	Uropoietic system
6**	Reproductive system
7**	Musculosceletal system

Table 3
Classification results based on 6231 images from 81 categories of at least five entries

Global feature	Similarity measure	1-NN (%)	5-NN (%)	Within 5 (%)	Within 10 (%)
Edge structure	Mahalanobis	17.46	21.78	40.06	89.17
DCT-based texture	Mahalanobis	40.80	43.94	60.82	92.33
Texture (CASTELLI)	Mahalanobis	39.51	42.29	61.27	93.36
Texture (TAMURA)	Jensen-Shannon	66.10	65.99	80.16	96.47
Re-scaled 8×8	Euclidian, EDM	70.92	70.69	82.54	96.79
	Covariance, CCF, $d=1$	70.84	72.59	84.45	97.59
Re-scaled 16×16	Euclidian, EDM	71.88	70.47	82.60	97.03
	Covariance, CCF, $d=2$	75.86	75.73	86.45	97.62
	Tangent, TDM	72.88	71.83	82.94	96.15
Re-scaled 24×24	Euclidian, EDM	71.79	70.33	82.51	96.95
	CCF, $d=3$	76.07	76.31	86.62	97.72
	Tangent, TDM	72.40	71.45	82.76	96.32
Re-scaled 32×32	Euclidian, EDM	71.58	70.18	82.31	96.89
	CCF, $d=4$	76.06	76.42	86.60	97.71
	Tangent, TDM	72.38	71.58	82.60	96.39
Down-scaled $h=32$	IDM	82.30	80.71	90.11	97.03
TAMURA & IDM	Parallel combination	85.48	85.36	92.97	95.25

The correctness is given for one-nearest-neighbor (1-NN) and five-nearest-neighbor (5-NN) classifiers. The frequency that the correct class occurs at least once within a set of k best responses is displayed for $k=5$ (within 5) and $k=10$ (within 10).

images and 5756 images from 54 categories) are provided in [36]. However, the figures obtained do not significantly differ from those presented in Table 3.

6. Discussion

Obviously, the recognition rates obtained by scaled representations outperform all global texture measures omitting any local information. Note that this result corresponds to previous investigations [26]. Nevertheless, the experiments show that global texture features are very useful to improve the categorization accuracy within a combined classifier, since their decision for each sample is less correlated with the decision made by the classifiers based on scaled representations.

The classifier combination improves the results because the two single classifiers evaluate different aspects of an image, Tamura texture features are global whereas the IDM keeps local pixel neighborhood information. The classifier results are therefore rather uncorrelated and allow to compensate errors from the single classifiers. With respect to combined classifier results, error rates of 15% remain (Table 3). In other words, 905 out of the 6231 images are misclassified. This is due to several reasons:

- The visual appearance of images in some categories still varies largely. This holds also for images coded with an identical IRMA code. For instance, Fig. 2 illustrates the high intra-category variability. All radiographs are coded with plain radiography, coronal posterioanterior direction, body region abdomen, and the musculoskeletal biosystem (1121-120-800-700). By grouping sparse categories into larger supersets, this variability is further increased: For example, from the 905 misclassified images, 124 images belong to a category with five or less samples when referring to the fully detailed code. This includes 43 images that lie in an isolated category with respect to the complete IRMA code.

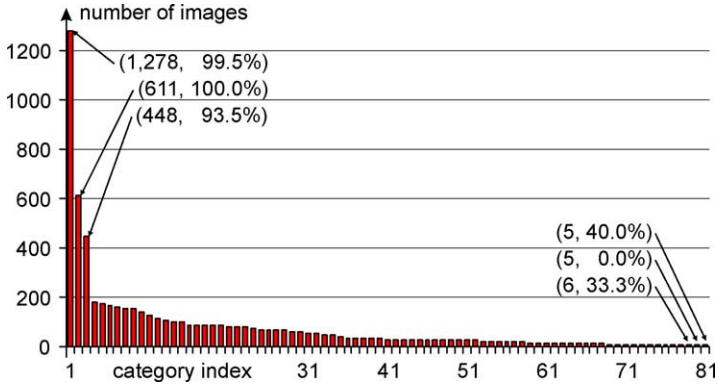

Fig. 3. Different sample sizes. The arrows annotate (number of samples, correctness of classification). The correctness for categories with many samples is significantly larger as compared to small sample sizes.

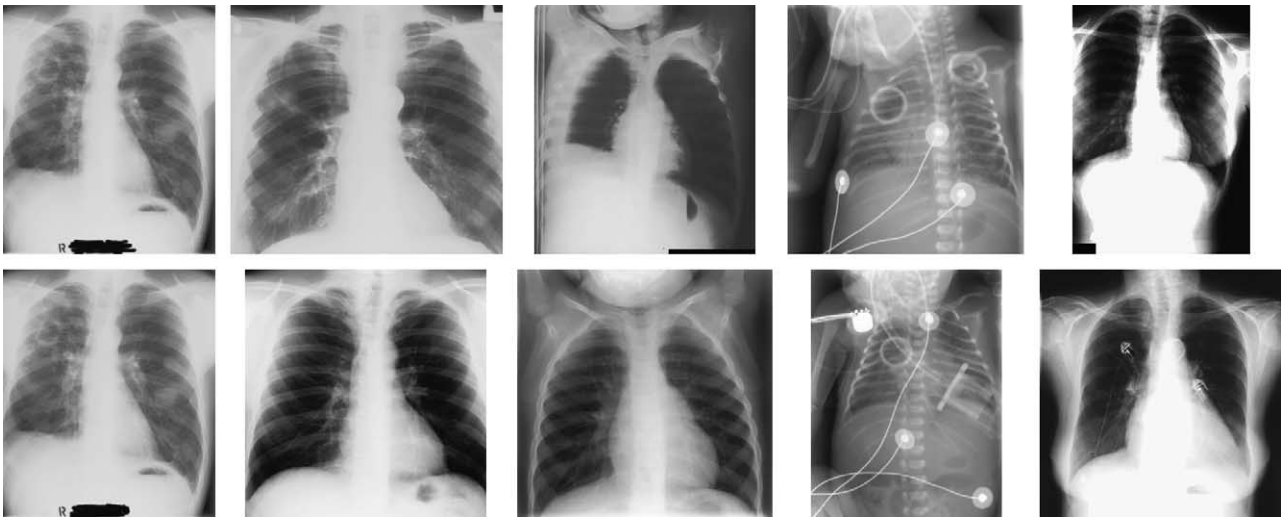

Fig. 4. The intra-class variability for large categories is compensated by their size: arbitrarily selected images from 11**-1**-50*-0** (upper row) and their respective nearest neighbors (lower row), which are also all from 11**-1**-50*-0**.

- As seen in Fig. 3, the categories differ considerably in size. In general, the recognition rates among the categories are very inhomogeneous. Almost all large categories have a recognition rate significantly above the overall rate of 85% whereas images from small classes are frequently misclassified. This shows that a sufficient number of representatives must be contained in the reference data. To come closer to this requirement, reference labelling of images is still in process. Note that large categories contain enough references to allow reliable recognition, even when pathologies or other alterations are present. Fig. 4 shows images selected arbitrarily from the largest category containing thoraces from coronal view.
- Although some categories differ in IRMA code, the images have a similar appearance. Fig. 5 illustrates this problem for mammographies that were acquired in craniocaudal and oblique orientation. Inspecting the confusion matrix reveals other cases such as finger vs. toe, upper arm vs. upper leg, or different projections of the cervical spine.

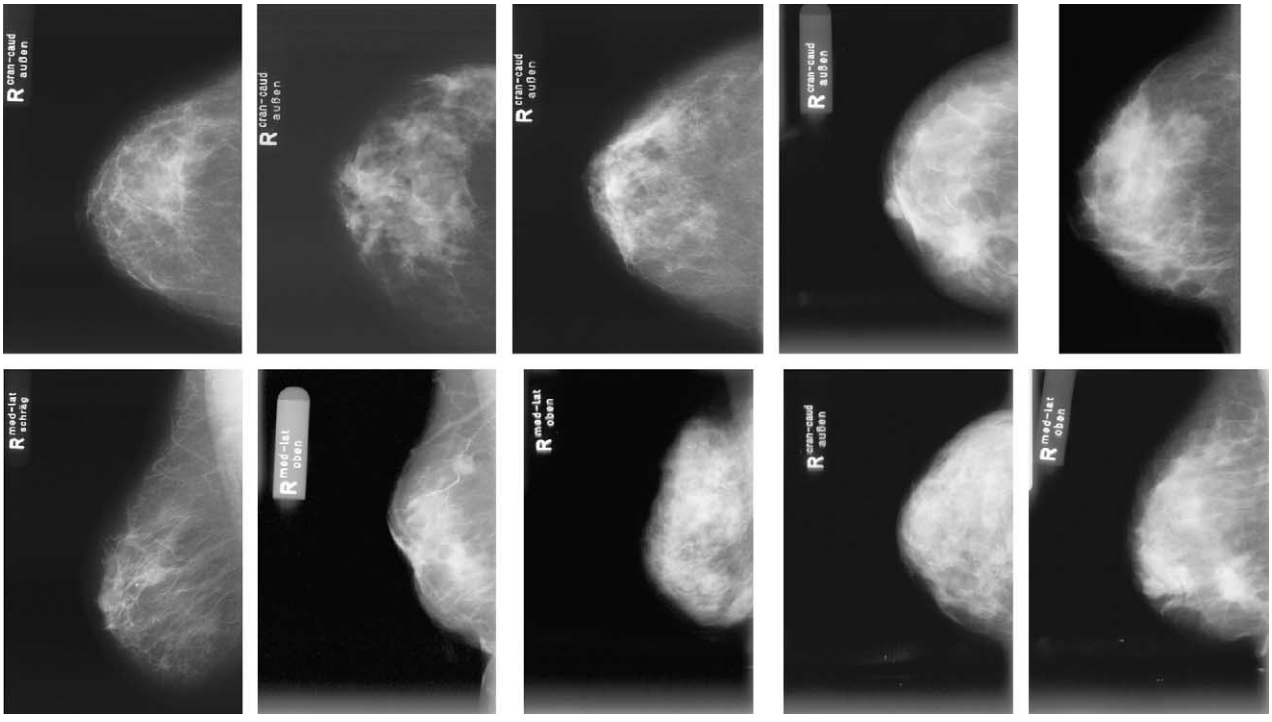

Fig. 5. Inter-class similarity. All mammographies in axial/craniocaudal view are coded 11**-3**-61*-6** (upper row), while other/oblique orientation is IRMA-coded 11**-4**-61*-6** (lower row).

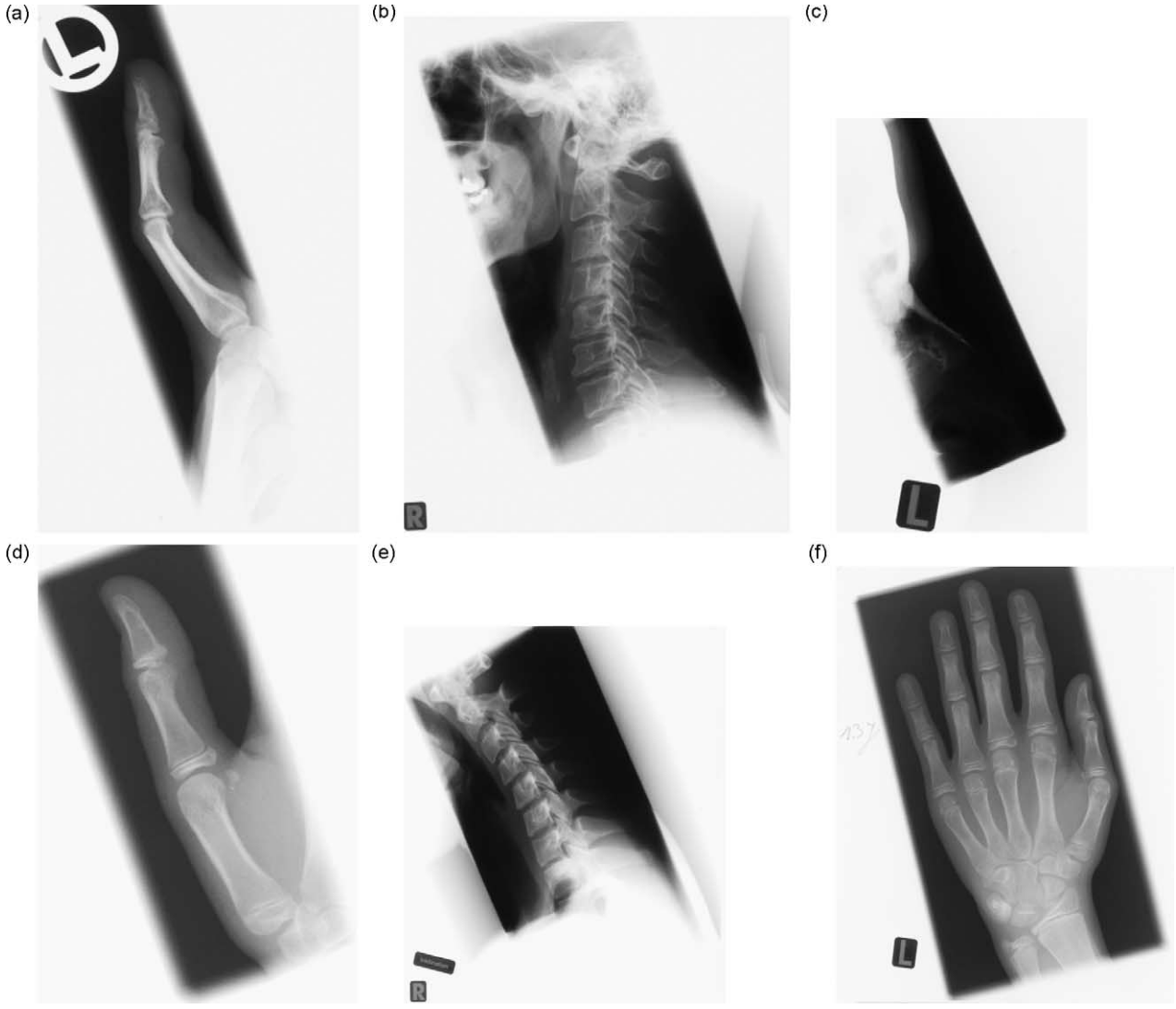

Fig. 6. Misclassification resulting from collimation field interference. The misclassified image is displayed on the left. Only the third nearest neighbor is from the correct class.

- The presence of collimation fields influences the feature extraction process and the similarity computations (Fig. 6). Especially the classifiers based on scaled representations are influenced by these areas since they produce significant contributions to distance calculation when comparing background (all-white or all-black) to image pixels. However, a preprocessing step which identifies and masks out collimation fields can be added to avoid this problem [37].

In the course of our experiments, the 1-NN seems to be the better choice for EDM and IDM, while for the remaining settings, the 5-NN led to better results or no significant difference are observed. The problem of determining the best setting for k is well known and general rules are hard to construct. The problem is reflected in our results, as even on the same data differences can be observed. More importantly, the hierarchical IRMA code employed to describe the image content allows to investigate the results at an arbitrary level of detail. This will help to develop a hierarchical classifier scheme in the future. For instance, a second classifier stage can be designed to distinguish different views of mammographies or fingers and toes, which frequently are confused based on a simple classification using global features. In addition, such schemes allow to incorporate a-priori knowledge on high semantical levels.

In conclusion, this work presents an extensive evaluation of automatic categorization using global features on a medical image corpus obtained from clinical routine. Even for a large number of 81 categories, a correctness rate of 82.3% was obtained using a single classifier based on scaled representations of the images and a similarity measure that is robust to local image deformations. The categorization rate is improved to 85.5% when a parallel combination of single classifiers based on scaled representations and global texture features is used. Considering image categorization

as initial step for medical CBIR, the correct image category should be within the five or ten nearest neighbors. In this case, recognition rates of 97.7% are obtained using re-scaled images with 24×24 pixels. Since further improvements of the automatic categorization may result from a hierarchical combination of classifiers, this correctness, which is based on global low-level image features only, can be regarded as sufficient for most applications in data mining. It allows to compile medical CBIR systems that are no longer restricted to a specific context. Likewise the IRMA approach, where image categorization is the very first step of image processing, the semantical knowledge about the image content enables appropriate selection of algorithms and their parameters for further image processing and analysis. Therefore, it is an important step to close the semantical gap of currently available systems.

7. Summary

This paper presents a comparative evaluation of methods for automatic categorization of medical images. Automatic categorization is a first step towards the use of data mining methods on medical image databases and it is obviously necessary for medical applications such as cased-based reasoning. Methods of content-based image access are applied taking into account the special properties of medical images.

Existing systems and features used in medical applications are briefly reviewed. Most of the systems are either applicable to a very limited task only or they strongly rely on alphanumerical descriptions or annotations that have to be created manually. Contrarily, the approach presented here is applicable to any type of medical imagery and not limited to a narrow set of tasks, since the automatical categorization allows for appropriate selection and parameterization of successive image processing steps.

So far, published approaches for automatic categorization are designed for a small number of categories, i.e. not more than 10 different classes. For instance, the separation of frontal and lateral views of chest radiographs has been frequently discussed in literature. The systems proposed are able to solve this two-class problem with a correctness up to 99.9%. They are based on global features, which means that a relatively small number of numerical values is used to describe the entire image.

However, medical images usually render some otherwise very successful discriminative features for images like color inapplicable. Therefore, texture and structure descriptors as well as down-scaled representations are evaluated as feature types using a nearest neighbor classifier and the automatic combination of classifier results. However, we still focus on global features, where the entire medical image is represented with less than 1024 numerical values.

Experiments for evaluation are carried out on a corpus of 6335 images selected arbitrarily from clinical routine. A reference categorization of the images is encoded using a multi-axial, mono-hierarchical coding scheme. This categorization was done by experienced radiologists familiar with the code. The hierarchy of the code allows the analysis of the automatic categorization performance (depending on the features and the classifier used) at various levels of differentiation. Experiments are done for 54 and 57 categories or 70 and 81 categories using radiographs only or all images, respectively. In particular, the experiments based on 6231 images from all kind of modalities, which were separated into 81 classes with at least 5 samples per class are analyzed. A maximum classification accuracy of 85% is obtained using a simple nearest neighbor classifier. Accuracy is increased to 98% if the correct category is only required to be within the ten best matches, which is sufficient for most applications in content-based image retrieval.

In conclusion, this work presents an extensive evaluation of different methods for automatic categorization of medical images. It is shown that the presented approaches are promising to offer new possibilities for content-based access to medical images as an accuracy of 98% within the ten best matches is sufficient for most applications. Thus, content-based image retrieval systems that are no longer limited to a special context are becoming possible.

Acknowledgements

This work is part of image retrieval in medical applications (IRMA), a research project funded by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG), grants Le 1108/4-1, Le 1108/4-2.

References

[1] Smeulders AWM, Worring M, Santini S, Gupta A, Jain R. Content-based image retrieval at the end of the early years. IEEE Trans Pattern Anal Machine Intell 2000;22(12):1349–80.
[2] Hand D, Manila H, Smyth P. Principles of data mining. Cambridge, MA: MIT Press; 2001.
[3] Niblack W, Barber R, Equitz W, Flickner M, Yanker P, Ashley J. The QBIC-project-Querying images by content using color, texture and shape. Proc SPIE 1993;1908:173–87.
[4] Flickner M, Sawhney H, Niblack W, Ashley J, Huang Q, Dom B, et al. Query by image and video content—The QBIC system. IEEE Comput 1995;28(9):23–32.
[5] Müller H, Michoux N, Bandon D, Geissbuhler A, A review of content-based image retrieval systems in medical applications. Clinical benefits and future directions. Int J Med Inform 2004;73:1–23.
[6] Tagare HD, Jaffe CC, Duncan J. Medical image databases—a content-based retrieval approach. J Am Med Informat Assoc 1997;4:184–98.
[7] Güld MO, Kohnen M, Schubert H, Wein BB, Lehman TM. Quality of DICOM header information for image categorization. Proc SPIE 2002;4685:280–7.
[8] Lehmann TM, Wein BB, Dahmen J, Bredno J, Vogelsang F, Kohnen M. Content-based image retrieval in medical applications-A novel multi-step approach. Proc SPIE 2000;3972:312–20.

[9] Long LR, Thoma GR. Landmarking and feature localization in spine x-rays. J Electronic Imaging 2001;10(4):939–56.
[10] Petrakis GM. Design and evaluation of spatial similarity approaches for image retrieval. Image Vision Comput 2002;20(1):59–76.
[11] Lehmann TM, Güld MO, Thies C, Fischer B, Spitzer K, Keysers D, Ney H, Kohnen M, Schubert H, Wein BB, Content-based image retrieval in medical applications. Meth Informat Med 2004;43(4):354–61.
[12] Love HJ, Antipov I, Hersh W, Smith CA, Mailhot M. Automated semantic indexing of imaging reports to support retrieval of medical images in the multimedia electronic medical record. Meth Inform Med 1999;38:303–7.
[13] Orphanoudakis SC, Chornaki C, Kostomanolakis S. I^2C—a system for the indexing, storage and retrieval of medical images by content. Med Informatics 1994;19(2):109–22.
[14] El-Kwae YEA, Xu H, Kabuka MR. Content-based retrieval in picture archiving and communication systems. IEEE Trans Knowledge Data Eng 2000;13(2):70–81.
[15] McG. Squire D, Müller W, Müller H, Raki J, Content-based query of image databases-Inspirations from text retrieval-Inverted files, Frequency-based weights and relevance feedback. Proceeding Scandinavian Conference on Image Analysis, Kangerlussuaq, Greenland; 1999. p. 143–9.
[16] Chu WW, Hsu CC, Cardenas AF, Taira RK. Knowledge-based image retrieval with spatial and temporal constructs. IEEE Trans Knowledge Data Eng 1998;10:872–88.
[17] Petrakis E, Faloutsos C, Lin KID. Imagemap—an image indexing method based on spatial similarity. IEEE Trans Knowledge Data Eng 2002;14(5):979–87.
[18] Wang JZ, Li J, Wiederhold G. SIMPLIcity-Semantics-sensitive integrated matching for picture libraries. IEEE Trans Pattern Anal Machine Intell 2001;23(9):947–63.
[19] Carson C, Belongie S, Greenspan H, Malik J. Blobworld-Image segmentation using expectation-maximization and its application to image querying. IEEE Trans Pattern Anal Machine Intell 2002;24(8): 1026–38.
[20] Barnard K, Duygulu P, Forsyth D. Modeling the statistics of image features and associated text. Proceedings document recognition and retrieval; 2002.
[21] Jain AK, Duin RPW, Mao J. Statistical pattern recognition—a review. IEEE Trans Pattern Anal Machine Intell 2000;22(1):4–36.
[22] Pietka E, Huang HK. Orientation correction for chest images. J Digital Imaging 1992;5(3):185–9.
[23] Boone JM, Seshagiri S, Steiner RM. Recognition of chest radiograph orientation for picture archiving and communications systems display using neural networks. J Digital Imaging 1992;5(3):190–3.
[24] Lehmann TM, Güld MO, Keysers D, Schubert H, Kohnen M, Wein BB. Determining the view position of chest radiographs. J Digital Imaging 2003;16(3):280–91.
[25] Pinhas A, Greenspan H. A continuous and probabilistic framework for medical image representation and categorization. Proc SPIE. 2003;5371:230–8.
[26] Keysers D, Dahmen J, Ney H, Wein BB, Lehmann TM. Statistical framework for model-based image retrieval in medical applications. J Electronic Imaging 2003;12(1):59–68.
[27] Lehmann TM, Schubert H, Keysers D, Kohnen M, Wein BB. The IRMA code for unique classification of medical images. Proc SPIE 2003;5033:109–17.
[28] Haralick RM, Shanmugam, Dinstein I. Textural features for image classification. IEEE Trans Syst, Man, Cybernetics 1973;SMC-3:610–21.
[29] Tamura H, Mori S, Yamawaki T. Textural features corresponding to visual perception. IEEE Trans Syst, Man, Cybernetics 1978;SMC-8(6):460–72.
[30] Puzicha J, Rubner Y, Tomasi C, Buhmann J. Empirical evaluation of dissimilarity measures for color and texture. Proc Int Conf Comput Vision 1999;2:1165–73.
[31] Castelli V, Bergman LD, Kontoyiannis I, Li CS, Robinson JT, Turek JJ. Progressive search and retrieval in large image archives. IBM J Res Dev 1998;42(2):253–68.
[32] Ngo CW, Pong TC, Chin RT. Exploiting image indexing techniques in DCT domain. Proceedings IAPR International workshop on multimedia information analysis and retrieval 1998;196–206.
[33] Zhou XS, Huang TS. Edge-based structural features for content-based image retrieval. Pattern Recogn Lett 2001;22(5):457–68.
[34] Simard PY, LeCun YA, Denker JS. Efficient pattern recognition using a new transformation distance. In: Hanson S, Cowan J, Giles J, editors. Advanced Neural Information Process System, vol. 5. San Mateo, CA: Morgan Kaufmann; 1993.
[35] Keysers D, Gollan C, Ney H. Classification of Medical Images using Non-linear Distortion Models. Proceedings Bildverarbeitung für die Medizin. Berlin: Springer; 2004 in press.
[36] Güld MO, Keysers D, Deselaers T, Leisten M, Schubert H, Ney H, Lehmann TM. Comparison of global features for categorization of medical images. Proc SPIE 2004;5371:211–22.
[37] Lehmann TM, Goudarzi S, Linnenbrügger NI, Keysers D, Wein BB. Automatic localization and delineation of collimation fields in digital and film-based radiographs. Proc SPIE 2002;4684(2):1215–23.

Thomas M. Lehman received the masters degree in electrical engineering (School of Engineering) and the PhD degree (summa cum laude) in computer science (School of Science), and the habilitation in medical informatics (School of Medicine) from the Aachen University of Technology (RWTH), Aachen, Germany, in 1992, 1998 and 2004, respectively. In 1992, he was research scientist at the Faculty of Electrical Engineering, RWTH Aachen. Since 1992, he has been with the Department of Medical Informatics, Medical Faculty RWTH Aachen, where he currently heads the Division of Medical Image Processing at the associated professor level (Privatdozent). He co-authored a textbook on image processing for the medical sciences (Springer-Verlag, Berlin, Germany, 1997) and co-edited the Handbook of Medical Informatics (Hanser Verlag, Munich, Germany, 2002). His research interests are discrete realizations of continuous image transforms, medical image processing applied to quantitative measurements for computer-assisted diagnoses and content-based image retrieval from large medical databases. Dr Lehmann received the DAGM-Preis'93. The award from the German Association for Pattern Recognition was given for his work on automatic strabometry using Hough transform and covariance filtering. In 1998, he received the Borcher's Medal from the RWTH Aachen for his work on medical image registration and interpolation. He is Chairman of the German Workshop on Medical Processing and Vice-President of the working group Medical Image Processing within the German Society of Medical Informatics, Biometry and Epidemiology (GMDS). He is Chairman of the IEEE Joint Chapter Engineering in Medicine and Biology (IEEE German Section). He is member of the Institute of Electrical and Electronics Engineers (IEEE), the International Association of Dentomaxillofacial Radiology (IADMFR), and the Society of Photo-Optical Instrumentations Engineering (SPIE). He serves on the International Editorial Board of Dentomaxillofacial Radiology.

Mark Oliver Güld received the Diploma degree in computer science from RWTH Aachen in 2001. Since then, he works as a PhD student at the Institute of Medical Informatics at the RWTH Aachen University Hospital. His main research topic is the implementation of a distributed image processing platform for image retrieval.

Thomas Deselaers received the Diploma degree in computer science (with honors) in 2004 from RWTH Aachen University, Germany. From June 2001 to March 2003, he was a student researcher at the Department of Computer Science of RWTH Aachen University. In September and October 2002, he was a visiting student researcher at the Instituto Tecnológico de Informática at the Unversidad Politécnica de Valencia, Spain. Since March 2004, he is a Research Assistant with the Department of Computer Science of the RWTH Aachen University. His research interests are in content-based image retrieval, analysis and classification of complex scenes, data mining, and pattern recognition.

Daniel Keysers received the Diploma degree in computer science (with honors) from the RWTH Aachen University, Germany, in 2000. Since then, he has been a PhD student and research assistant at the Department of Computer Science of the RWTH, where he currently is the head of the image processing and object recognition group at the Chair of Computer Science VI. His research interests include statistical modeling for pattern recognition, invariance in image object recognition and computer vision, and (medical) image retrieval.

Henning Schubert studied human medical sciences at the Aachen University of Technology (RWTH), Germany. He is a Board certified Radiologist and works for the Department of Diagnostic Radiology, University Hospital Aachen. His main interests are image processing, computer vision and reasoning, and organizational systems (picture archiving and communication systems, radiology information systems).

Klaus Spitzer received the masters degree in mathematics and the PhD degree (magna cum laude) in mathematics from the Friedrich-Wilhelms-University, Bonn, Germany, in 1983 and 1985, respectively. In 1983, he earned the MD degree (summa cum laude) from the Friedrich-Wilhelms-University, Bonn, Germany. From 1983 to 1993, he was a physician and neurologist at the Department of Neurology, University Hospital of Hamburg, Germany. He was a professor of neurology at the University of Hamburg from 1992 to 1994 and an associate professor of medical informatics from 1993 to 1995 at the University of Heidelberg, Germany. Since 1995, he has been a professor of medical informatics and the chair of the Institute of Medical Informatics, Aachen University of Technology (RWTH), Aachen, Germany. His research interests are knowledge-based systems in medicine, computer-based training, digital records, hospital information systems, and management of IT-systems.

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 work 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, NJ. In July 1993, he joined RWTH Aachen (University of Technology), Germany, as a professor for 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 modelling, 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 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).