Comparison of Global Features for Categorization of Medical Images

Mark Oliver Güld[a], Daniel Keysers[b], Thomas Deselaers[b], Marcel Leisten[c], Henning Schubert[c], Hermann Ney[b], and Thomas M. Lehmann[a]

[a]Department of Medical Informatics, Medical Faculty
[b]Chair of Computer Science VI, Department of Computer Science
[c]Department of Diagnostic Radiology, Medical Faculty
Aachen University of Technology, Aachen, Germany

ABSTRACT

We present an evaluation of methods for the automatic categorization of medical images. The properties of medical images render some otherwise very successful discriminate features for images (e.g. color) inapplicable. Therefore, we evaluate several feature types: texture, structure, and down-scaled representations. The classification is done using a nearest neighbor classifier with various distance measures as well as the automatic combination of classifier results. A corpus of 6,335 images selected arbitrarily from the clinical routine was encoded using a multi-axial, mono-hierarchical code. The reference categorization was done by experienced radiologists familiar with the code. The code's hierarchy allows the analysis of the automatic categorization performance (depending on the features and the classifier used) at different levels of differentiation. Experiments were done for 54 and 57 categories or 70 and 81 categories focussing on radiographs only or for all images, respectively. A maximum classification accuracy of 86% was obtained using the winner-takes-all rule and a one nearest neighbor classifier. Accuracy is increased to 93% and 95% if the correct category is only required to be within the 5 or 10 best matches, respectively. In this case, the best rate of 98% is obtained. This is sufficient for most applications in content-based image retrieval.

Keywords: pattern recognition, medical image retrieval, feature extraction

1. INTRODUCTION

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. Usually, the categorization is done manually by the physician or radiologist during the routine documentation. In modern picture archiving and communication systems (PACS), digital modalities are connected via the digital imaging and communications in medicine (DICOM) protocol. 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. Unfortunately, this information cannot always be considered reliable.[1] Especially in applications of digital radiology, such as computer-aided diagnosis (CAD), or content-based image retrieval (CBIR), the image category is of major importance for subsequent processing steps.[2] Here, image categorization allows context-specific selection of appropriate filters or algorithmic parameters.

In general, automatic categorization as a mapping of images into their classes involves three basic principles[3]: (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.

Corresponding author: Dr. Thomas Lehmann, Department of Medical Informatics, Aachen University of Technology, Pauwelsstr. 30, D – 52057 Aachen, Germany, email: lehmann@computer.org; web: http://irma-project.org/lehmann, phone: +49 241 80-88793.

Medical Imaging 2004: PACS and Imaging Informatics,
edited by Osman M. Ratib, H. K. Huang, Proceedings of SPIE Vol. 5371
(SPIE, Bellingham, WA, 2004) · 1605-7422/04/$15 · doi: 10.1117/12.535914

Considering representation, a set of global features is extracted from each of the images and combined to a feature vector. Here, the term 'global feature' means that only a small number of numerical values are used to describe the entire image. For instance, most CBIR-systems that are designed for a general purpose, do not perform retrieval on the object level. Instead, they perform a classification task using a small set of features to describe the image content. An example for such a system is the query by image content (QBIC) system from IBM designed to browse internet databases.[4] Basically, three major types of features are used for image descriptions: color, contour, and texture, with color being the most successfully used feature in general purpose CBIR.[5] With respect to medical image data, color features are mostly inapplicable. Contour descriptors can only be applied successfully if the extraction of a closed contour is possible in all images of the corpus, e.g. for images containing isolated objects and a homogeneous background. However, typical properties of radiographs, e.g. summation effects and noise, render the automatic extraction of contours extremely difficult, even if the context is well known.

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 is differentiated by means of digital image processing.[6,7] For this two-class experiment, the error rates are below 1%.[8] In a recent investigation, PINHAS and GREENSPAN report error rates below 1% for categorization of 851 medical images into eight classes.[9] In a paper of KEYSERS et al., six classes are defined according to the body part examined from 1,617 images and an error rate of 8% is reported.[10]

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. Within the project for content-based image retrieval in medical applications (IRMA, http://irma-project.org), a detailed classification scheme has been developed to encode medical images according to their content.[11] The four axes of the IRMA code assess the imaging technique and modality (T-axis, 4 levels of detail), the relative direction of the imaging device and the patient (D-axis, 3 levels of detail), the anatomic body part that is examined (A-axis, 3 levels of detail), and the biological system being under investigation (B-axis, 3 levels of detail). Thus, each image encoding has the form TTTT-DDD-AAA-BBB, with presently 797 unique entities available on the four axes. With respect to plain radiography, about 10,000 images have been taken randomly from clinical routine 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. For instance, if the IRMA code is compressed to only 2, 1, 2, and 1 code positions at the T, D, A, and B axis, respectively, about 80 used categories remain. However, this is still much more that the two or eight classes that have been analyzed so far.

This paper aims at analyzing 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. Therefore, the paper is organized as follows. In the next section, we briefly describe the image corpus, its manual reference coding, the global features which are applied for categorization, and the similarity measures that can be used for classification. In Section 3, results are presented according to an exhaustive data analysis. The measured effects are discussed in detail (Sec. 4) before the conclusion is drawn (Sec. 5).

2. METHODS

2.1. The image corpus

At the date of our experiments, the corpus contains 6,335 images that have been taken arbitrarily from clinical routine at the Aachen University Hospital. Mostly, secondary digitized images from plain radiography (5,839 images) but also images from other modalities, e.g. computed tomography and ultrasound imaging were collected (Table 1). All images have been categorized by an experienced radiologist according to the IRMA code.[11] 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 codes. Additionally, we use thresholds for the minimum number of images in each category and drop all images from categories below the threshold. This results in 6,231 images from

technique and modality modality	number of images	
	absolute	relative
plain radiography	5,839	92.17 %
fluoroscopy	104	1.64 %
angiography	101	1.59 %
computed tomography (CT)	212	3.35 %
ultrasound, b-mode	17	0.27 %
ultrasound, Doppler-mode	1	0.02 %
magnetic resonance imaging (MRI)	61	0.96 %

Table 1. Distribution of the images among the T-axis.

81 categories and 6,155 images from 70 categories when using a minimum of five or ten images per category, respectively.

Since the IRMA project is currently focused on plain radiography, images from other modalities are so far rather under-represented (Table 1). Therefore, we also investigate classification results for radiographs only. Using the same thresholds, we obtain 5,776 images from 57 categories and 5,756 images from 54 categories for a minimum of five and ten images per category, respectively. Table 2 shows the category distribution among the radiographs in the image corpus.

2.2. 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. Considering texture, a wide range of features has been proposed in the literature. Based on several experiments, those features being most suitable to distinguish medical images have been chosen.

2.2.1. Texture features proposed by Tamura

Based on fundamental work of Haralick and coworkers,[12] Tamura et al. suggested coarseness, contrast, and directionality to describe an image's texture properties.[13] These features are computed on a per-pixel basis. Therefore, we collect the values into a three-dimensional histogram ($6 \times 8 \times 8 = 384$ bins) and use the Jensen-Shannon-divergence to measure the similarity between two histograms.[14] To make the texture properties comparable, all images were scaled into an identical size of 256×256 pixels, ignoring the initial aspect ratio.

2.2.2. Texture features proposed by Castelli

Castelli et al. used various texture features to describe image properties.[15] 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 from scaled images of fixed size (256×256 pixels).

2.2.3. Texture features proposed by Ngo

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.[16] 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. Again, the extraction was performed on identically sized images (256×256 pixels).

IRMA code	number of images		IRMA code	number of images	
	absolute	relative		absolute	relative
11⋆⋆-1⋆⋆-20⋆-7⋆⋆	174	3.01 %	11⋆⋆-2⋆⋆-22⋆-7⋆⋆	15	0.26 %
11⋆⋆-1⋆⋆-21⋆-7⋆⋆	43	0.74 %	11⋆⋆-2⋆⋆-23⋆-7⋆⋆	179	3.10 %
11⋆⋆-1⋆⋆-31⋆-7⋆⋆	152	2.63 %	11⋆⋆-2⋆⋆-31⋆-7⋆⋆	157	2.72 %
11⋆⋆-1⋆⋆-32⋆-7⋆⋆	82	1.42 %	11⋆⋆-2⋆⋆-32⋆-7⋆⋆	89	1.54 %
11⋆⋆-1⋆⋆-33⋆-7⋆⋆	139	2.41 %	11⋆⋆-2⋆⋆-33⋆-7⋆⋆	165	2.86 %
11⋆⋆-1⋆⋆-41⋆-7⋆⋆	448	7.76 %	11⋆⋆-2⋆⋆-41⋆-7⋆⋆	63	1.09 %
11⋆⋆-1⋆⋆-42⋆-7⋆⋆	25	0.43 %	11⋆⋆-2⋆⋆-42⋆-7⋆⋆	24	0.42 %
11⋆⋆-1⋆⋆-43⋆-7⋆⋆	21	0.36 %	11⋆⋆-2⋆⋆-43⋆-7⋆⋆	34	0.59 %
11⋆⋆-1⋆⋆-44⋆-7⋆⋆	28	0.48 %	11⋆⋆-2⋆⋆-44⋆-7⋆⋆	28	0.48 %
11⋆⋆-1⋆⋆-45⋆-7⋆⋆	65	1.13 %	11⋆⋆-2⋆⋆-45⋆-7⋆⋆	5	0.09 %
11⋆⋆-1⋆⋆-46⋆-7⋆⋆	99	1.71 %	11⋆⋆-2⋆⋆-46⋆-7⋆⋆	36	0.62 %
11⋆⋆-1⋆⋆-50⋆-0⋆⋆	1,278	22.13 %	11⋆⋆-2⋆⋆-50⋆-0⋆⋆	611	10.58 %
11⋆⋆-1⋆⋆-51⋆-7⋆⋆	54	0.93 %	11⋆⋆-2⋆⋆-91⋆-7⋆⋆	29	0.50 %
11⋆⋆-1⋆⋆-70⋆-4⋆⋆	116	2.01 %	11⋆⋆-2⋆⋆-92⋆-7⋆⋆	64	1.11 %
11⋆⋆-1⋆⋆-70⋆-5⋆⋆	14	0.24 %	11⋆⋆-2⋆⋆-93⋆-7⋆⋆	20	0.35 %
11⋆⋆-1⋆⋆-71⋆-4⋆⋆	15	0.26 %	11⋆⋆-2⋆⋆-94⋆-7⋆⋆	105	1.82 %
11⋆⋆-1⋆⋆-72⋆-4⋆⋆	10	0.17 %	11⋆⋆-2⋆⋆-95⋆-7⋆⋆	36	0.62 %
11⋆⋆-1⋆⋆-73⋆-4⋆⋆	8	0.14 %	11⋆⋆-2⋆⋆-96⋆-7⋆⋆	44	0.76 %
11⋆⋆-1⋆⋆-80⋆-2⋆⋆	7	0.12 %	11⋆⋆-3⋆⋆-61⋆-6⋆⋆	80	1.39 %
11⋆⋆-1⋆⋆-80⋆-7⋆⋆	124	2.15 %	11⋆⋆-3⋆⋆-62⋆-6⋆⋆	82	1.42 %
11⋆⋆-1⋆⋆-91⋆-7⋆⋆	86	1.49 %	11⋆⋆-3⋆⋆-94⋆-7⋆⋆	60	1.04 %
11⋆⋆-1⋆⋆-92⋆-7⋆⋆	64	1.11 %	11⋆⋆-4⋆⋆-21⋆-7⋆⋆	155	2.68 %
11⋆⋆-1⋆⋆-93⋆-2⋆⋆	24	0.42 %	11⋆⋆-4⋆⋆-23⋆-7⋆⋆	28	0.48 %
11⋆⋆-1⋆⋆-93⋆-7⋆⋆	15	0.26 %	11⋆⋆-4⋆⋆-31⋆-7⋆⋆	31	0.54 %
11⋆⋆-1⋆⋆-94⋆-7⋆⋆	101	1.75 %	11⋆⋆-4⋆⋆-41⋆-7⋆⋆	85	1.47 %
11⋆⋆-1⋆⋆-95⋆-2⋆⋆	10	0.17 %	11⋆⋆-4⋆⋆-61⋆-6⋆⋆	86	1.49 %
11⋆⋆-1⋆⋆-95⋆-7⋆⋆	45	0.78 %	11⋆⋆-4⋆⋆-62⋆-6⋆⋆	87	1.51 %
11⋆⋆-1⋆⋆-96⋆-7⋆⋆	75	1.30 %	11⋆⋆-4⋆⋆-91⋆-7⋆⋆	29	0.50 %
11⋆⋆-2⋆⋆-21⋆-7⋆⋆	27	0.47 %	⋆⋆⋆⋆-⋆⋆⋆-⋆⋆⋆-⋆⋆⋆	5,776	100.00 %

Table 2. The image corpus and the categories used in the experiments (radiographs only, min. category size=5).

2.2.4. Image structure proposed by Zhou and Huang

In 2001, Zhou and Huang proposed an algorithm to capture properties of edges within an image.[17] A water-filling process is applied to the binarized gradient image. Canny's edge detector is used to determine the gradient. The three parameters, 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' suggestion, we used the filling time, fork count, and loop count, both counts computed for a global and a per-edge-segment maximum. Again, the feature extraction was performed on images with 256×256 pixels.

2.2.5. Down-scaled representations

In previous work, down-scaled images have been used successfully as feature vectors.[18] To obtain vectors of identical size $h \times h$, the images are again scaled ignoring their aspect ratio. For this type of feature, several similarity measures can be applied besides general Euclidian distance. This allows to integrate a-priori knowledge about class-invariant transformations, making the classifier more robust, especially when dealing with few training examples.

Adopted from signal processing, the cross-correlation function (CCF)

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

returns the maximum correlation for a pair of images $r(x,y)$ and $s(x,y)$ over a selected warp range, i.e. the two-dimensional translations over d pixels are performed explicitly. In Equation (1), $\bar{r}$ and $\bar{s}$ denote the pixel-wise mean gray value of the reference and sample image r and s, respectively. For our experiments, we used $d = \lfloor h/8 \rfloor$. The correlation function also normalizes image brightness, which is another common cause of variability found in medical images.

Alternatively, the image distortion model (IDM) allows local displacements for each pair of corresponding pixels compared within the distance measure.[19] 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 information is evaluated. 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

$$D_{\mathrm{IDM}}(r,s) = \sum_{x=1}^{X}\sum_{y=1}^{Y} \min_{|x'|,|y'|\leq W_1} \left\{ \sum_{|x''|,|y''|\leq W_2} ||r(x+x'+x'',y+y'+y'')-s(x+x'',y+y'')||_2 \right\} \quad (2)$$

The results are improved if the image gradient is used instead of the intensity values. For our experiment, we used $W_1 = 2$ (5×5 pixel-sized search window for corresponding pixels) and $W_2 = 1$ (3×3 pixels of local context). The images were scaled to a fixed height of 32 pixels keeping their original aspect ratio.

2.3. Automatic Classifiers

A 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 measure. 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.

2.4. Classifier combination

Classifier combination can be grouped into three main categories[3]: parallel, serial (like a sieve), and hierarchical (comparable to a tree). We use parallel classifier combination, since it is an easy way to post-process existing results obtained from the single classifiers. 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 was used to optimize the weighing coefficients, which were then applied to combine the results for the full image corpus.[20]

2.5. Evaluation

Based on the image corpus, exhaustive experiments were carried out using the leaving-one-out scheme for evaluation. Each time, one image is used as the test image and the remaining images as references. Then, the mean categorization rate over all iterations 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). Since the IRMA-concept proposes to pursue the most likely categories for each unknown image for further content abstraction,[2] it was also investigated whether the correct category occurs among the first five or ten neighbors. This estimates how many hypotheses must be kept for subsequent processing steps.

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	70.92 %	70.69 %	82.54 %	96.79 %
	CCF, $d = 1$	70.84 %	72.59 %	84.45 %	97.59 %
Re-scaled 16×16	Euclidian	71.88 %	70.47 %	82.60 %	97.03 %
	CCF, $d = 2$	75.86 %	75.73 %	86.45 %	97.62 %
Re-scaled 24×24	Euclidian	71.79 %	70.33 %	82.51 %	96.95 %
	CCF, $d = 3$	76.07 %	76.31 %	86.62 %	97.72 %
Re-scaled 32×32	Euclidian	71.58 %	70.18 %	82.31 %	96.89 %
	CCF, $d = 4$	76.06 %	76.42 %	86.60 %	97.71 %
Down-scaled $h = 32$	IDM	82.30 %	80.71 %	90.11 %	97.03 %

Table 3. Classification results using single classifiers (all images, min. category size=5: 6,231 images, 81 categories).

Global feature	Similarity measure	1-NN	5-NN	within 5	within 10
Edge structure	Mahalanobis	17.94 %	22.16 %	40.65 %	89.80 %
DCT-based texture	Mahalanobis	41.12 %	44.52 %	61.17 %	92.64 %
Texture (CASTELLI)	Mahalanobis	39.97 %	42.92 %	61.71 %	93.66 %
Texture (TAMURA)	Jensen-Shannon	66.42 %	66.30 %	80.44 %	96.57 %
Re-scaled 8×8	Euclidian	71.16 %	71.08 %	82.84 %	96.93 %
	CCF, $d = 1$	71.06 %	73.06 %	84.70 %	97.86 %
Re-scaled 16×16	Euclidian	72.07 %	70.90 %	82.92 %	97.16 %
	CCF, $d = 2$	76.15 %	76.08 %	86.71 %	97.87 %
Re-scaled 24×24	Euclidian	71.97 %	70.77 %	82.81 %	97.09 %
	CCF, $d = 3$	76.31 %	76.65 %	86.95 %	97.95 %
Re-scaled 32×32	Euclidian	71.78 %	70.64 %	82.62 %	97.03 %
	CCF, $d = 4$	76.34 %	76.82 %	86.90 %	97.94 %
Down-scaled $h = 32$	IDM	82.57 %	81.12 %	90.33 %	97.16 %

Table 4. Classification results using single classifiers (all images, min. category size=10: 6,155 images, 70 categories).

3. RESULTS

3.1. Single Classifiers

According to the experiment setup with two minimum category sizes and the two image sets, four tables of results are obtained. Tables 3 and 4 show the results for all images using a minimum of five and 10 images per category, respectively, Tables 5 and 6 contain the results for all radiographs, i.e. the IRMA code on the T-axis is fixed to 11⋆⋆.

The best recognition rates are obtained using all images and a threshold of ten images per category (70 categories). For a threshold of five images per category (81 categories), the rates do not drop significantly. When only radiographs are considered, the recognition rates drop by about 0.5%-2% in absolute numbers.

The feature describing properties of the edge structure performs worst in all experiments and does not exceed 22.5% recognition rate. Texture features proposed by CASTELLI and the features based on NGO's approach perform on a similar level. 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% for all images and 42.2% for all radiographs (regarding five images per category in each case) resulted. The histograms based on TAMURA's texture features yielded the best results among the features proposed for general-purpose image retrieval: 66% correctness for all

Global feature	Similarity measure	1-NN	5-NN	within 5	within 10
Edge structure	Mahalanobis	17.76 %	22.13 %	40.36 %	89.84 %
DCT-based texture	Mahalanobis	38.63 %	42.15 %	59.22 %	92.14 %
Texture (CASTELLI)	Mahalanobis	37.73 %	40.72 %	60.15 %	93.33 %
Texture (TAMURA)	Jensen-Shannon	64.46 %	64.61 %	79.29 %	96.38 %
Re-scaled 8×8	Euclidian	70.12 %	70.26 %	82.39 %	96.88 %
	CCF, $d = 1$	69.75 %	72.00 %	84.25 %	97.85 %
Re-scaled 16×16	Euclidian	70.98 %	70.14 %	82.44 %	97.11 %
	CCF, $d = 2$	75.16 %	75.31 %	86.27 %	97.87 %
Re-scaled 24×24	Euclidian	70.91 %	70.12 %	82.24 %	97.04 %
	CCF, $d = 3$	75.33 %	75.95 %	86.50 %	97.94 %
Re-scaled 32×32	Euclidian	70.62 %	69.94 %	82.13 %	96.97 %
	CCF, $d = 4$	75.38 %	76.09 %	86.41 %	97.91 %
Down-scaled $h = 32$	IDM	81.79 %	80.56 %	89.96 %	97.11 %

Table 5. Classification results using single classifiers (radiographs, min. category size=5: 5,776 images, 57 categories).

Global feature	Similarity measure	1-NN	5-NN	within 5	within 10
Edge structure	Mahalanobis	17.91 %	22.26 %	40.58 %	90.10 %
DCT-based texture	Mahalanobis	38.77 %	42.38 %	59.41 %	92.35 %
Texture (CASTELLI)	Mahalanobis	37.86 %	40.91 %	60.32 %	93.52 %
Texture (TAMURA)	Jensen-Shannon	64.46 %	64.61 %	79.29 %	96.38 %
Re-scaled 8×8	Euclidian	70.24 %	70.40 %	82.57 %	96.98 %
	CCF, $d = 1$	69.87 %	72.22 %	84.36 %	97.95 %
Re-scaled 16×16	Euclidian	71.07 %	70.26 %	82.59 %	97.19 %
	CCF, $d = 2$	75.33 %	75.43 %	86.36 %	97.97 %
Re-scaled 24×24	Euclidian	71.00 %	70.24 %	82.38 %	97.12 %
	CCF, $d = 3$	75.47 %	76.08 %	86.61 %	98.04 %
Re-scaled 32×32	Euclidian	70.71 %	70.10 %	82.28 %	97.05 %
	CCF, $d = 4$	75.54 %	76.23 %	86.52 %	98.00 %
Down-scaled $h = 32$	IDM	81.93 %	80.68 %	90.10 %	97.19 %

Table 6. Classification results using single classifiers (radiographs, min. category size=10: 5,756 images, 54 categories).

images and 64.6% for all radiographs (using a minimum category size of five). 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 8x8 images, the most basic approach. Euclidian distance yields around 72% recognition rate in each experiment. The correlation function, which adds robustness with respect to translations and intensity changes, yields 76% for all images and 75.4% for the radiographs (regarding a minimum of five images per category). On very small images, it performs worse than Euclidian distance but the additional image information from larger representations improves the accuracy, while Euclidian distance starts to be negatively affected by small variations in translation for representations larger than 16x16. IDM for representations scaled to a fixed height of 32 pixels yields the best results: 82.3% for all images and 81.8% for all radiographs using a minimum of five images per category. Contrary to the texture features, the best results among the scaled representations are obtained using 1-NN (IDM). Only the classifier using CCF benefits from taking into account more than the nearest neighbor.

With respect to routine applications of CBIR in medicine, it is interesting whether the correct class is within a fixed number of best responses. Taking into account the first five neighbors, the IDM contains in 90% of the cases one or more members of the correct category for each experiment setup. Considering the ten nearest neighbors, CCF performs best using 81 categories and an image representation of 24×24 pixels and contains for

Corpus	1-NN	5-NN	within 5	within 10
All images, min. 5 images/category	85.48 %	85.36 %	92.97 %	95.25 %
All images, min. 10 images/category	85.69 %	85.65 %	93.19 %	95.43 %
Radiographs, min. 5 images/category	85.01 %	85.01 %	92.78 %	95.19 %
Radiographs, min. 10 images/category	85.15 %	85.18 %	92.95 %	95.33 %

Table 7. Classification results using classifier combination.

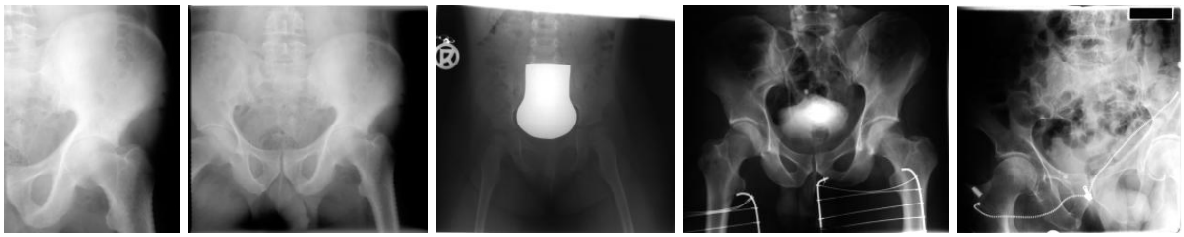

Figure 1. Intra-class variability. All radiographs are coded identically (IRMA 1121-120-800-700).

97.7% of all images a member of the correct class.

All experiments were done using ordinary PC hardware. The scaled representations need little preprocessing time (several seconds for high-resolution images), while the texture features are more expensive at extraction time. The classification itself can be executed within milliseconds, since the distance measures are simple and the data vectors are very small (9 to 284 components). IDM, the most complex classifier was used with pre-filtering, i.e. the IDM is applied only to the nearest 100 neighbors that have been obtained by Euclidian distance. This let the classification time drop below one second per sample.

3.2. Classifier combination

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) was evaluated (Table 7). Regarding a minimum of five images per category, the recognition rate is significantly improved from 81.8% to 85.0% and from 82.3% to 85.5% for the radiographs only and for all images, respectively. The rate of neighborhoods containing one or more members from the correct category is also increased, e.g. from 90.0% to 92.8% for radiographs and a minimum category size of five. However, the rate for the first ten neighbors does not. In fact, it is lower than for both single classifier results.

4. DISCUSSION

Usually, increasing the number of categories decreases the classification quality. Comparing the results for all images and the sub-group of radiographs, only slight differences are observed. This effect can be explained by several reasons. (i) Images from other modalities differ significantly in appearance from the radiographs. (ii) With respect to CT and MRI, adjacent slices form dense clusters in feature space. However, it is important that for all classes a sufficient number of images covering the intra-class variability is included in the reference data.

Obviously, the recognition rates obtained by scaled representations outperform all global texture measures omitting any local information. Note that this results corresponds to previous investigations.[10] 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.

With respect to combined classifier results, error rates of 15% remain (Table 7). Considering the radiographs and a minimum number of five images in each category, 866 out of 5,776 images are misclassified. This is due to several reasons:

category code	number of images	errors	recognition rate
11⋆⋆-1⋆⋆-50⋆-0⋆⋆	1,278	7	99.45 %
11⋆⋆-2⋆⋆-50⋆-0⋆⋆	611	1	99.84 %
11⋆⋆-1⋆⋆-41⋆-7⋆⋆	448	28	93.75 %
11⋆⋆-2⋆⋆-23⋆-7⋆⋆	179	5	97.21 %
11⋆⋆-1⋆⋆-20⋆-7⋆⋆	174	2	98.85 %
	...		
11⋆⋆-1⋆⋆-95⋆-2⋆⋆	10	13	0.00 %
11⋆⋆-1⋆⋆-72⋆-4⋆⋆	10	28	10.00 %
11⋆⋆-1⋆⋆-73⋆-4⋆⋆	8	30	75.00 %
11⋆⋆-1⋆⋆-80⋆-2⋆⋆	7	6	14.29 %
11⋆⋆-2⋆⋆-45⋆-7⋆⋆	5	9	0.00 %

Table 8. The largest and smallest classes: A sufficient number of images is required for high recognition rates.

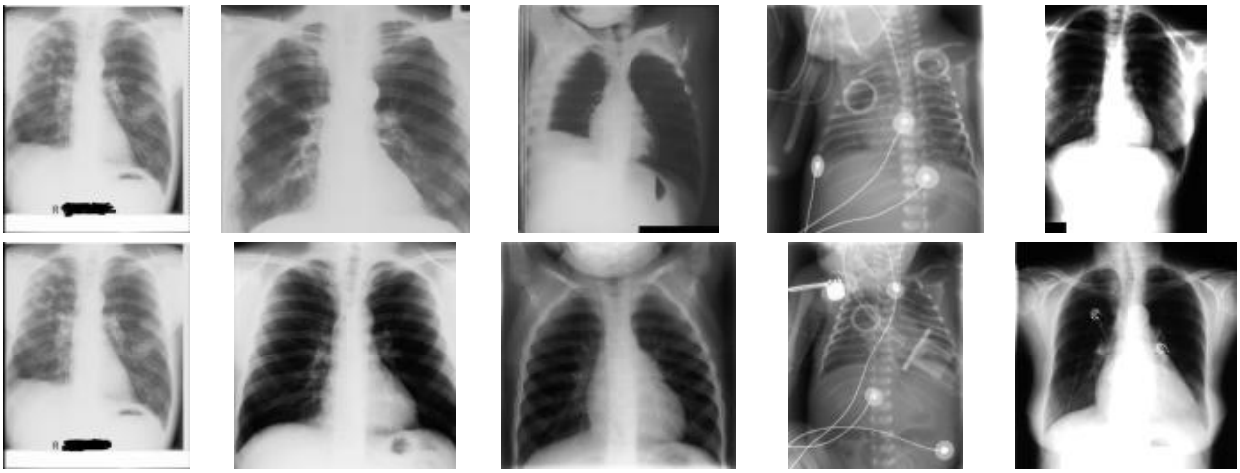

Figure 2. The intra-class variability for large categories is compensated by their size: arbitrarily selected images from 11⋆⋆-1⋆⋆-50⋆-0⋆⋆(upper row) and their respective nearest neighbors (lower row).

- The visual appearance of images in some categories still varies largely. This holds also for images coded all with an identical IRMA-code. For instance, Figure 1 illustrates the high intra-category variability. All radiographs are coded with plain radiography, coronal pa 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 866 misclassified images, 42 images lie in an isolated category when referring to the fully detailed code.

- As seen in Table 2, 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 85,01% whereas images from small classes are frequently misclassified (Table 8). Again, this shows that a sufficient number of representatives must be contained in the reference data. Therefore, 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. Figure 2 shows images selected arbitrarily from the largest category containing thoraces from coronal view.

- Another reason for classification failures are classes with different IRMA-code but similar appearance. Figure 3 illustrates this problem. While the images in the upper row are projected in axial/craniocaudal orientation (code 310 for the D-axis), the examples shown in the lower row are acquired in other/oblique direction (code 410 for the D-axis). Inspecting the confusion matrix discovers other cases such as finger

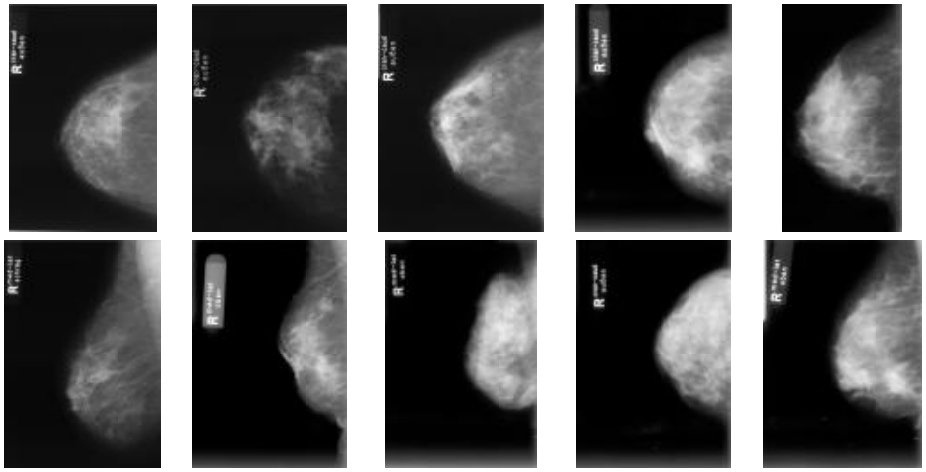

Figure 3. Example images from categories 11⋆⋆-3⋆⋆-61⋆-6⋆⋆(upper row) and 11⋆⋆-4⋆⋆-61⋆-6⋆⋆(lower row).

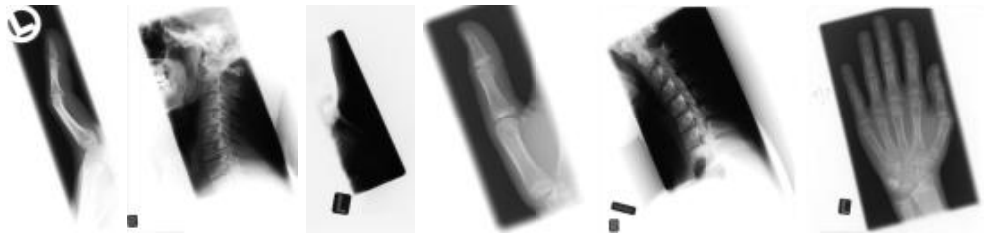

Figure 4. Misclassification resulting from collimation field interference.

vs. toe, upper arm vs. upper leg, or different projections of the cervical spine.

- The presence of collimation fields influences the feature extraction process and the similarity computations. Figure 4 illustrates a misclassification that results from the presence of collimation fields. The misclassified image is displayed on the left. Only the third neighbor is from the correct class. Especially the classifiers based on scaled representations are effected by these areas since they produce significant contributions to distance calculation when comparing background (all-white or all-black) to image pixels. A preprocessing step which identifies and masks out collimation fields must be added to avoid this problem. Several algorithms have been proposed to address this issue.[21]

In the course of our experiments, we used the two choices $k = 1$ and $k = 5$ in the k-NN classifier. As can be observed, the 1-NN seems to be the better choice for the Euclidian distance and the IDM, while for the remaining settings the 5-NN led to better results or no significant difference could be 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.

The hierarchical IRMA-code employed to describe the image content allows to investigate the results at an arbitrary level of detail. This will help in future to develop a hierarchical classifier scheme. Such schemes allow to incorporate a-priori knowledge on semantical levels.

5. 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 categorization 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 can be further 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 image retrieval based on local features, the correct image category should be within the five or ten nearest neighbors. In this case, recognition rates of 98% is obtained using the Re-scaled images with 24×24 pixels (Table 5). This result, which is obtained from global image features, can be regarded as sufficient for most applications. However, further improvements of the automatic categorization may result from a hierarchical combination of classifiers.

6. ACKNOWLEDGMENT

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

REFERENCES

1. Güld MO, Kohnen M, Schubert H, Wein BB, Lehman TM: Quality of DICOM Header Information for Image Categorization. Proceedings SPIE 4685: 280-287, 2002.
2. Lehmann TM, Wein BB, Dahmen J, Bredno J, Vogelsang F, Kohnen M: Content-Based Image Retrieval in Medical Applications A Novel Multi-Step Approach. Proceedings SPIE 3972: 312-320, 2000.
3. Jain AK, Duin RPW, Mao J: Statistical Pattern Recognition: A Review. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(1): 4-36, 2000.
4. Flickner M, Sawhney H, Niblack W, Ashley J, Huang Q, Dom B, Gorkani M, Hafner J, Lee D, Petkovic D, Steele D, Yanker P: Query by image and video content – The QBIC system. IEEE Computer 28(9): 23-32, 1995.
5. Smeulders AWM, Worring M, Santini S, Gupta A, Jain R: Content-Based Image Retrieval at the End of the Early Years. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(12): 1349-1380, 2000.
6. Pietka E, Huang HK (1992) Orientation correction for chest images. Journal of Digital Imaging 5(3): 185-189, 1992.
7. Boone JM, Seshagiri S, Steiner RM: Recognition of chest radiograph orientation for picture archiving and communications systems display using neural networks. Journal of Digital Imaging 5(3): 190-193, 1992.
8. Lehmann TM, Güld MO, Keysers D, Schubert H, Kohnen M, Wein BB: Determining the view position of chest radiographs. Journal of Digital Imaging 16(3): 280-291, 2003.
9. Pinhas A, Greenspan H: A continuous and probabilistic framework for medical image representation and categorization. Proceedings SPIE Medical Imaging 2004, in press for this conference.
10. Keysers D, Dahmen J, Ney H, Wein BB, Lehmann TM: Statistical Framework for Model-based Image Retrieval in Medical Applications. Journal of Electronic Imaging, 12(1): 59-68, 2003.
11. Lehmann TM, Schubert H, Keysers D, Kohnen M, Wein BB: The IRMA code for unique classification of medical images. Proceedings SPIE 5033: 109-117, 2003.
12. Haralick RM, Shanmugam, Dinstein I: Textural features for image classification. IEEE Transactions on System, Man, and Cybernetics SMC-3: 610-621, 1973.
13. Tamura H, Mori S, Yamawaki T: Textural Features Corresponding to Visual Perception. IEEE Transactions on Systems, Man, and Cybernetics; SMC-8(6), 460-472, 1978.
14. Puzicha J, Rubner Y, Tomasi C, Buhmann J: Empirical Evaluation of Dissimilarity Measures for Color and Texture. Procs. International Conference on Computer Vision, Vol. 2, 1165-1173, 1999.
15. Castelli V, Bergman LD, Kontoyiannis I, Li CS, Robinson JT, Turek JJ: Progressive Search and Retrieval in Large Image Archives. IBM Journal of Research and Development 42(2): 253-268, 1998.
16. Ngo CW, Pong TC, Chin RT: Exploiting Image Indexing Techniques in DCT Domain. IAPR International Workshop on Multimedia Information Analysis and Retrieval: 196-206, 1998.
17. Zhou XS, Huang TS: Edge-based structural features for content-based image retrieval. Pattern Recognition Letters 22(5): 457-468, 2001.
18. Lehmann TM, Güld MO, Keysers D, Schubert H, Wenning A, Wein BB: Automatic detection of the view position of chest radiographs. Proceedings SPIE 5032(3): 1275-1282, 2003.

19. Keysers D, Gollan C, Ney H: Classification of Medical Images using Non-linear Distortion Models. Proceedings BVM 2004, Bildverarbeitung fr die Medizin 2004, Springer-Verlag, Berlin, in press.
20. Deselaers T: Features for image retrieval. Diploma Thesis, Chair of Computer Science VI, RWTH Aachen University, 2004.
21. Lehmann TM, Goudarzi S, Linnenbrügger NI, Keysers D, Wein BB: Automatic Localization and Delineation of Collimation Fields in Digital and Film-Based Radiographs. Procs SPIE 2002; 4684(2): 1215-1223.