Similarity of Medical Images Computed from Global Feature Vectors for Content-Based Retrieval

Thomas M. Lehmann[1], Mark O. Güld[1], Daniel Keysers[2], Thomas Deselaers[2], Henning Schubert[3], Berthold Wein[3], and Klaus Spitzer[1]

[1] Department of Medical Informatics, Aachen University of Technology (RWTH)
Pauwelsstr. 30, D - 52057 Aachen, Germany
{tlehmann, mgueld, kspitzer}@mi.rwth-aachen.de
http://irma-project.org
[2] Chair of Computer Science VI, Aachen University of Technology (RWTH)
Ahornstr. 55, D - 52056 Aachen, Germany
{keysers, deselaers}@informatik.rwth-aachen.de
[3] Department of Diagnostic Radiology, Aachen University of Technology (RWTH)
Pauwelsstr. 30, D - 52057 Aachen, Germany
{schubert, wein}@rad.rwth-aachen.de

Abstract. Global features describe the image content by a small number of numerical values, which are usually combined into a vector of less than 1,024 components. Since color is not present in most medical images, grey-scale and texture features are analyzed in order to distinguish medical imagery from various modalities. The reference data is collected arbitrarily from radiological routine. Therefore, all anatomical regions and biological systems are present and all images have been captured in various directions. The ground truth is established by manually reference coding with respect to a mono-hierarchical unambiguous coding scheme. Based on 6,335 images, experiments are performed for 54 and 57 categories or 70 and 81 categories focusing on radiographs only or considering all images, respectively. A maximum classification accuracy of 86% was obtained using the winner-takes-all rule and a one nearest neighbor classifier. If the correct category is only required to be within the 5 or 10 best matches, we yield a best rate of 98% using normalized cross correlation of small image icons.

1 Introduction

For efficient computation of image similarity, 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 is used to describe the entire image. An example for such a system is the query by image content (QBIC) system from IBM which is designed to browse internet databases [1]. Basically, three major types of features are used for image descriptions: color, contour, and texture. It has been shown that color is the most successfully used feature in general purpose CBIR systems [2].

With respect to medical imagery, color features are mostly inapplicable. Furthermore, contour descriptors can only be applied successfully if the extraction of a

M.Gh. Negoita et al. (Eds.): KES 2004, LNAI 3214, pp. 989–995, 2004.

closed contour is reliable 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 effect and noise, render the automatic extraction of contours extremely difficult, even if the context is well known. Consequently, texture features are applied for content-based access to medical images. In particular, global texture features have been used for categorization of medical images.

The strong relationship between image retrieval and image categorization has been pointed out by Liu et al. [3]. So far, automatic categorization 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 automatically distinguished by means of digital image processing [4, 5]. For this two-class experiment, the error rates are below 1% [6]. In a recent investigation, error rates below 1% are reported for categorization of 851 medical images into eight classes [7]. In other work, six classes are defined according to the body part examined from 1,617 images and an error rate of 8% is reported [8]. However, such a low number of classes is not suitable for content-based image retrieval applied to evidence-based medicine or case-based reasoning. Here, the image category must be determined in much more detail as a first step of processing [9]. This paper analyses the use of global features for automatic image categorization into a large number of classes.

2 Materials and Methods

2.1 Establishing the Ground Truth

To compare approaches for automatic classification and similarity computing of medical images, a ground truth or gold standard is required. Referring to [10], a gold standard must be (i) reliable, i.e. the generation or capturing of test images for evaluation must follow an exactly determined and reproducible protocol, (ii) equivalent, i.e. the image material or relationships considered within an algorithmic reference standard must compare to real-life data with respect to structure, noise, or other parameters of importance, and (iii) independent, i.e. any reference standard must rely on a different procedure than that to be evaluated, or on other images or image modalities than those to be evaluated.

Equivalence is guaranteed when images are collected from clinical routine. Within the project for content-based image retrieval in medical applications (IRMA, http://irma-project.org), about 10,000 two-dimensional images have been taken randomly from clinical routine. Independence is obtained if leaving-one-out experiments are applied for which all images are classified by human experts. In order to guarantee reliability, manual references must be independent of the expert labeling the images. Therefore, 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.

Reference coding of the IRMA database resulted in more than 400 used codes. In contrast to other coding schemes, the IRMA code is mono-hierarchical, 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. Table 1 shows the different sets of radiographs used in our experiments [12].

Table 1. Taking advantage of the hierarchical structure of IRMA code, different data sets were compiled for leaving-one-out experiments

Data set number	Template of IRMA code	Minimum # of samples	Total # of images	Total # of categories
1	TT**-D**-AA*-B**	5	6,231	81
2	TT**-D**-AA*-B**	10	6,115	70
3	11**-D**-AA*-B**	5	5,776	57
4	11**-D**-AA*-B**	10	5,756	54

2.2 Selecting Global Features and Similarity Measures

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. Table 2 shows the texture features and their references. The first four methods refer to rather rigid texture and structure measures, while the latter four also cope with global or local image deformations.

Table 2. Global image features and similarity measures included in this study

Number	Type	Similarity / Distance	Authors	References
1	texture	Jensen-Shannon	Tamura et al.	[13]
2	texture	Mahalanobis	Castelli et al.	[14]
3	texture	Mahalanobis	Ngo et al.	[15]
4	structure	Mahalanobis	Zhou & Huang	[16]
5	scaled	Euclidean	Lehmann et al.	[6]
6	scaled	Cross Covariance	Lehmann et al.	[6]
7	scaled	Tangent Distance	Keysers et al.	[8]
8	scaled	Image Distortion Model	Keysers et al.	[17]

Using Euclidean distance, cross covariance, or the tangent distance measure, the original images were scaled down to h x h pixels, $h \in \{32, 24, 16, 8\}$, regardless of the initial aspect ratio. Regarding the image distortion model, the aspect ratio is maintained and a size of 32 x b or b x 32 pixels, $b \leq 32$, is chosen for portray and landscape formats, respectively.

2.3 Selecting Classifiers and Classifier Combination

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, $k = 1$ is chosen. Data based on $k = 5$ is published elsewhere [12]. Note that this is a simple yet effective method, which is also useful to present classification results interactively.

Classifier combination can be grouped into three main categories [18]: (i) parallel, (ii) serial (like a sieve), and (iii) hierarchical (comparable to a tree). We used 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 a specific set of categories.

For parallel combination, the classifier results are first transformed into 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.

3 Results

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 those 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. Here, a best recognition rate of 40.8%, 41.1%, 38.6%, and 38.8% resulted for the test sets 1, 2, 3, and 4, respectively. The histograms based on Tamura's texture features yield the best results among the features proposed for general-purpose image retrieval: 66.1%, 66.4%, 64.5%, and 64.5%, respectively.

In general, the scaled representations perform better than all texture features examined. Even for the Euclidian distance on 8 x 8 pixel icons, which is the most basic approach on a feature vector of 64 components, the correctness is 70.9%, 71.2%, 70.1%, and 70.2% for the test sets 1, 2, 3, and 4, respectively. For $h = 24$, the normalized correlation function, which adds robustness with respect to translations and intensity changes, yields 76.1%, 76.3%, 75.3%, and 75.5%, respectively. 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 h = 16. The image distortion model outperforms all other methods yielding 82.3%, 82.6%, 81.8%, and 81.9%, respectively.

Normalized cross correlation and image distortion model acknowledge spatial variability within a local neighborhood while the texture features capture rather global image properties. Therefore, a combination of classifiers based on the image distortion model (best among scaled representations) and the texture features according to Tamura (best among global texture features) was evaluated. The resulting correctness yields 85.5%, 85.7%, 85.0%, and 85.2%, respectively.

With respect to routine applications of CBIR in medicine, it is interesting whether the correct class is within a fixed number of best responses, which will be displayed to the physician for decision making. Taking into account the first five neighbors, the cross correlation performs best based on icons of 24 x 24 pixels resulting in a correctness of 97.7%, 97.9%, 97.9%, and 98.0% for the test sets no. 1, 2, 3, and 4, respectively.

4 Discussion

In most applications of data mining and content-based image retrieval, a ground truth or gold standard is unavailable. Therefore, concepts such as precision are frequently used [19], which do not evaluate the total number of correct images within the database. In our experiments, based on the unambiguous IRMA code, a gold standard was established and results were compared by means of their actual correctness.

The reasons for remaining errors are manifold. For instance, all results were computed from relative distances and similarity measures. Applying the winner-takes-all rule does not consider the actual distance, which might be large for misclassified examples. In addition, the data collected from clinical routine is highly variant (Fig. 1). The considerable intra-class variability is further enlarged by hierarchically merging the IRMA categories. In addition, some categories differ in IRMA code but not in appearance. This low inter-category variability can be observed, for instance, comparing the craniocaudal and the oblique view of x-ray mammographies. However, the first is acquired in axial direction while the latter refers to other directions. Also, fingers and toes or different areas of the spine are difficult to distinguish. Here, a hierarchical classifier can be established with specialized features and decision rules which can be optimized to the particular task. Since global representations are used, shutters or collimation fields significantly alternate the computed images features, and, consequently, image similarity is decided based on the shape of the shutter but not on the image content within the collimation field. Automatic collimation field detection, as proposed by Wiemker et al. [20], may overcome this problem.

Another reason for misclassification results from the unequal number of reference images per category. Reflecting the frequency of x-ray examinations, plain chest radiography is the class with most samples. While in data set no. 4, a total of 1,278 (22.1%) and 611 (10.6%) images are frontal and lateral views of chest radiographs, which are coded by 11**-1**-50*-0** and 11**-2**-50*-0**, respectively, 51 of 54 categories come with less than 200 samples. Therefore, correctness depends on the number of reference samples available. The error rate for categories with a small number of references is significantly higher than that of a large number. For instance, frontal chest radiographs are correctly detected with a mean correctness of 99.5 % and 100 % using the parallel combination of 1-NN classifiers and tracking the classes within the five nearest neighbors, respectively. This is due to the sufficient number of samples covering the intra-class variability (Fig. 1).

In summary, the figures presented prove that global image features are suitable for content-based retrieval of medical images. However, the references used for nearest neighbor classification must cover the entire variety of image appearances.

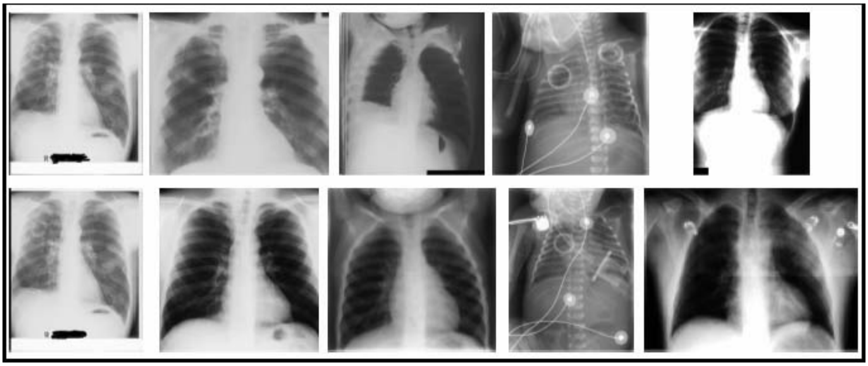

Fig. 1. The samples of high intra-class variance are taken from the IRMA category 11**-1**-50*-0**, chest radiographs in frontal view (upper row). The corresponding nearest neighbors (lower row) are all from the same category

5 Acknowledgement

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

References

1. 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 1995; 28(9): 23-32
2. 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 2000; 22(12): 1349-1380
3. Liu Y, Dellaert F, Rothfus WE: Classification driven semantic based medical image indexing and retrieval. Technical Report CMU-RI-TR-98-25, The Robotics Institute, Carnegie Mellon University, Pittsgurgh, PA, 1998
4. Pietka E, Huang HK (1992) Orientation correction for chest images. Journal of Digital Imaging 1992; 5(3): 185-189

5. 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 1992; 5(3): 190-193
6. Lehmann TM, Güld MO, Keysers D, Schubert H, Kohnen M, Wein BB: Determining the view position of chest radiographs. Journal of Digital Imaging 2003; 16(3): 280-291
7. Pinhas A, Greenspan H: A continuous and probabilistic framework for medical image representation and categorization. Proceedings SPIE Medical Imaging 2004, in press
8. Keysers D, Dahmen J, Ney H, Wein BB, Lehmann TM: Statistical framework for model-based image retrieval in medical applications. Journal of Electronic Imaging 2003; 12(1): 59-68
9. Müller H, Michoux N, Bandon D, Geissbuhler A: A review of content-based image retrieval systems in medical applications – Clinical benefits and future directions. International Journal of Medical Informatics 2004, in press
10. Lehmann TM: From plastic to gold: A unified classification scheme for reference standards in medical image processing. Proceedings SPIE 2002; 4684(3): 1819-1827
11. Lehmann TM, Schubert H, Keysers D, Kohnen M, Wein BB: The IRMA code for unique classification of medical images. Proceedings SPIE 2003; 5033: 109-117
12. Güld MO, Keysers D, Leisten M, Schubert H, Lehmann TM: Comparison of global features for categorization of medical images. Proceedings SPIE 2004; in press
13. Tamura H, Mori S, Yamawaki T: Textural features corresponding to visual perception. IEEE Transactions on Systems, Man, and Cybernetics 1978; SMC-8(6), 460-472
14. 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 1998 42(2): 253-268
15. Ngo CW, Pong TC, Chin RT: Exploiting image indexing techniques in DCT domain. IAPR International Workshop on Multimedia Information Analysis and Retrieval 1998; 196-206
16. Zhou XS, Huang TS: Edge-based structural features for content-based image retrieval. Pattern Recognition Letters 2001; 22(5): 457-468
17. Keysers D, Gollan C, Ney H: Classification of medical images using non-linear distortion models. Proceedings BVM 2004 (Bildverarbeitung für die Medizin), Springer-Verlag, Berlin, 2004; 366-370
18. Jain AK, Duin RPW, Mao J: Statistical pattern recognition: A review. IEEE Transactions on Pattern Analysis and Machine Intelligence 2000; 22(1): 4-36
19. Müller H, Müller W, McG Squire D, Marchand-Maillet S, Pun T: Performance evaluation in content-based image retrieval – Overview and proposals. Pattern Recognition Letters 2001; 22(5): 593-601
20. Wiemker R, Dippel S, Stahl M, Blaffert T, Mahlmeister U: Automated recognition of the collimation field in digital radiography images by maximization of the Laplace area integral. Proceedings SPIE 2000; 3979: 1555-1565