

The IRMA Reference Database and Its Use for Content-Based Image Retrieval in Medical Applications

Lehmann TM¹, Fischer B¹, Güld MO¹, Thies C¹, Keysers D², Deselaers T², Schubert H³, Wein BB³, and Spitzer K¹

¹Department of Medical Informatics, Aachen University of Technology, Aachen, Germany

²Chair of Computer Science VI, Aachen University of Technology, Aachen, Germany

³Departement Diagnostic Radiology, Aachen University of Technology, Aachen, Germany
lehmann@computer.org

Introduction

In medical image processing and analysis, a steadily growing demand on well defined references is inherent. Reference images are linked to pre-defined semantics and usually called ground truth or gold standard. In [1], it has been pointed out that a gold standard must be (i) reliable in itself, which means that manually drawn regions, manual selections, or manual classifications usually should not be referred to as gold standard since they may vary according to the non-reproducible interaction of humans, (ii) generated or obtained independently from the procedure to be evaluated, which means, for instance, that the same image should not be used both for training or parameter optimization of the algorithm and for its evaluation, and (iii) sufficiently reflecting the effect to be evaluated, which means, in particular, that a gold standard must reflect the manifold of appearances of the diagnostic imagery. With respect to the large variability of medical images having been acquired in clinical routine, a sufficient number of references is necessary. Most frequently, image processing algorithms developed for medical applications are evaluated on an insufficient number of images or based on images violating one or more criteria of a gold standard. Consequently, published algorithms of computer-assisted image analysis often lack in accuracy, robustness, or applicability.

With respect to the evaluation of computational methods of data mining or content-based access to medical images, an even larger set of gold standards is required. Beside the collection of the image data, a reliable labelling with the pre-defined semantics is required. In this paper, we describe the methods developed by the image retrieval in medical applications (IRMA, <http://irma-project.org>) project [2] to establish a reference database and its use to evaluate computer algorithms for automatic categorization of medical images.

Material and Methods

The IRMA Database. In order to build a sufficiently large database of reference images, diagnostic images have been selected arbitrarily from the routine records of the Department of Diagnostic Radiology at the University Hospital in Aachen, Germany. X-ray films were secondarily digitized and direct digital images were transferred using a DICOM interface. This procedure of image selection ensures that

the distribution of imaged body regions corresponds to their frequency in diagnostic routine.

The IRMA Reference Code. In order to establish a reliable ground truth, a coding scheme was developed that allows unique labelling of images without any inter- or intra-observer variability [3]. In contrast to existing coding schemes, such as the SNOMED vocabulary or the MeSH thesaurus, the IRMA code is strictly mono-hierarchical and unambiguous. For instance, the DICOM tag *body_part_examined* has valid entries *arm*, *hand*, or *extremity*, which results in ambiguous references. Consequently, the IRMA code ensures on all of its four axes, namely technique, direction, anatomy, and biosystem, that the reference coding assigned manually by experienced radiologists does not vary. Therefore, IRMA-coded images can be regarded as a gold standard for medical image categorization.

Computer-Assisted Reference Coding. Each image that is appended to the IRMA reference database is compared to existing images which have been labelled already. A web-based interface assists the radiologist by displaying the ten images from the database looking most similar to the query image and offering their IRMA code to be directly adopted. However, the entire coding scheme is also available in easy-to-use pull-down menus to allow encoding of images of such type that have not been included into the database before.

Evaluation of Automatic Coding. Leaving-one-out experiments have been carried out in order to evaluate the automatic categorization of images. Each image of the database was used once as a query and its IRMA code was compared to the nearest neighbour that has been determined by global texture features. By means of global features, the entire image is represented by a small number of numerical feature values, which are usually combined into a single feature vector. Beside texture features and rescaled images, parallel combinations of classifiers were also analysed.

Results

Currently, the IRMA database contains about 17,700 valid images, 10,746 of which are completely coded according to the IRMA code. At time of experiments, 6,335 images were used for evaluation. The IRMA code defines 345, 87, 173, and 192 entries on the technique, direction, anatomy, and biosystem axes, respectively. In total, this offers about one billion different categories. However, not all of the combinations are meaningful. In fact, the 6,335 images are distributed over 405 distinct codes. Taking advantage from the hierarchical structure of the IRMA code, several subsets have been compiled and used for evaluation: (a) 6,231 images of all modalities forming 81 categories with a minimum number of 5 samples per class, (b) 6,155 images of all modalities forming 70 categories with a minimum number of 10 samples per class, (c) 5,776 radiographs of 57 categories with a minimum of 5 samples per class, and (d) 5,756 radiographs of 54 categories with a minimum number of 10 samples per class. Based on the texture description proposed by Tamura et al. [3], the best categorization is obtained with a correctness of 66.10%, 66.42%, 64.42%, and 64.42% for the data sets (a), (b), (c), and (d), respectively. Operating most accurately on images rescaled according to their aspect ratio smaller than 32×32 images, the image distortion model [4] yields correctness of 82.30%, 82.57%, 81.79%, and 81.93%. A parallel combination of both methods improves the

results to 85.48%, 85.69%, 85.01%, and 85.15% for the data sets (a), (b), (c), and (d), respectively. Requiring the correct category to be within the 5 or 10 next neighbours, the normalized cross correlation based on down-sampled icons of 24 x 24 pixels disregarding the initial aspect ratio yields a correctness of 86.62% or 97.72%, 86.95% or 97.95%, 86.50% or 97.94%, and 86.61% or 98.04%, for the data sets (a), (b), (c), and (d), respectively.

Discussion and Conclusion

Neither the number of categories nor the number of images used in the experiments significantly affect the results. Hence, it can be concluded that the IRMA database in fact represents a gold standard for image categorization. The evaluation of automatic categorization procedures shows that a correctness of 85% is obtained by combining global texture features. Remaining errors can be explained from the partially high intra-class variability and the inter-class similarity. For instance, plain radiographs of fingers or toes look quite identical. In addition, collimation fields and shutter strongly affect the global texture features. Also, the varying frequency of samples in each category has a certain impact on the results. Classes with few samples produce significantly more errors than those with a large number. With respect to content-based image retrieval and data mining techniques, hypothesis can be tracked for further processing. Then, it is sufficient to find the correct one among the n next likely categories. Using the correlation of small image icons, the correctness increases to 98% for $n = 10$. Here, the feature vector used for global image representation is composed from less than 576 entries. In conclusion, IRMA technology allows automatic categorization of medical images with respect to their imaging modality and technical parameters, the system geometry of imaging device and patient, the body region examined, and the biological system under investigation, if a sufficient number of references is available. Hence, intelligent processing schemes can now be introduced that always select the optimal algorithm or parameterization for each image individually. The development and evaluation of this technology was based on a large database of gold standard images, which might also be useful for evaluation of other image processing methods in medicine.

Acknowledgement

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] Lehmann TM: From Plastic to Gold - A Unified Classification Scheme for Reference Standards in Medical Image Processing. Proceedings SPIE 2002; 4684(3): 1819-1827.
- [2] 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. Methods of Information in Medicine 2004; in press.
- [3] Lehmann TM, Schubert H, Keysers D, Kohnen M, Wein BB: The IRMA code for unique classification of medical images. Proceedings SPIE 2003; 5033: 440-451.
- [4] 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.
- [5] Keysers D, Gollan C, Ney H: Classification of Medical Images using Non-linear Distortion Models. Proceedings BVM 2004; in press.

