



Short Paper

The CLEF 2005 Automatic Medical Image Annotation Task

THOMAS DESELAERS

*Human Language Technology and Pattern Recognition, Computer Science Department, RWTH Aachen University,
Germany*

deselaers@cs.rwth-aachen.de

HENNING MÜLLER

Medical Informatics Service, Geneva University and Hospitals, Geneva Switzerland

henning.mueller@sim.hcuge.ch

PAUL CLOUGH

Department of Information Studies, Sheffield University, Sheffield, UK

p.d.clough@sheffield.ac.uk

HERMANN NEY

*Human Language Technology and Pattern Recognition, Computer Science Department, RWTH Aachen University,
Germany*

ney@cs.rwth-aachen.de

THOMAS M. LEHMANN

Department of Medical Informatics, Medical Faculty, RWTH Aachen University, Germany

lehmann@computer.org

Received June 6, 2006; Accepted August 15, 2006

First online version published in December, 2006

Abstract. In this paper, the automatic annotation task of the 2005 CLEF cross-language image retrieval campaign (ImageCLEF) is described. This paper focuses on the database used, the task setup, and the plans for further medical image annotation tasks in the context of ImageCLEF. Furthermore, a short summary of the results of 2005 is given. The automatic annotation task was added to ImageCLEF in 2005 and provides the first international evaluation of state-of-the-art methods for completely automatic annotation of medical images based on visual properties.

The aim of this task is to explore and promote the use of automatic annotation techniques to allow for extracting semantic information from little-annotated medical images. A database of 10.000 images was established and annotated by experienced physicians resulting in 57 classes, each with at least 10 images. Detailed analysis is done regarding the (i) image representation, (ii) classification method, and (iii) learning method. Based on the strong participation of the 2005 campaign, future benchmarks are planned.

Keywords: content-based image retrieval, medical image annotation, evaluation in computer vision

1. Introduction

Evaluating performance is a very important step in the development and investigation of new research methods. In speech recognition, machine translation and information retrieval, large-scale managed evaluation events are a common way to compare the performance of different systems. Examples are the NIST (National Institute of Standards and Technology) machine translation evaluation¹, the TC-STAR evaluation², the International Workshop on Spoken Language Translation (IWSLT) Evaluation³, and the NIST information retrieval evaluation campaign, called TREC⁴ (Text REtrieval Conference). In the field of image processing and recognition, evaluation is only recently becoming adopted: Benchathlon⁵ is an initiative for evaluating technologies including image filtering, content-based image retrieval (CBIR) and automatic description of images in large-scale image databases. However, to date no evaluation campaign has been carried out. ImageEVAL⁶ has done a preliminary test evaluation and is preparing for an official evaluation campaign. TRECVID⁷ is an evaluation campaign for video retrieval in the context of the Text REtrieval Conference (TREC)⁸ and has been organizing annual benchmarking events since 2001. In the context of the PASCAL network of excellence⁹, evaluation campaigns for object classification, detection, and segmentation methods were carried out in March 2005 (Everingham et al., 2006) and in April/May 2006.

The Cross Language Evaluation Forum (CLEF)¹⁰ aims at supporting global digital library applications by developing an infrastructure for testing, tuning, and evaluation of information retrieval systems. In particular, CLEF creates test-suites of reusable data which can be employed by system developers to benchmark their systems. In contrast to TREC, CLEF focusses on multi-lingual and more recently on multi-modal aspects of information retrieval. ImageCLEF¹¹ began as a pilot experiment in 2003 with a bilingual ad hoc retrieval task consisting of a database of images with accompanying texts in one language. The database was searched using textual queries in other languages. ImageCLEF 2003 attracted just four participants with approaches using text-based retrieval and query enhancement techniques. In 2004, a medical and an interactive retrieval task were added to ImageCLEF. The medical task used a set of images with associated medical case notes and was primarily offered as a query-by-visual-example (QBVE) retrieval task as search tasks supplied by the organisers contained only images and not text. However, participants could involve text in subsequent retrieval iterations and combine both image processing and text-based retrieval methods. ImageCLEF 2004 attracted strong participation from 18 research groups across the world, demonstrating the need for such an evaluation event. In 2005, the herein described medi-

cal automatic annotation task was added to ImageCLEF and again, participation increased. A total of 36 groups registered for ImageCLEF, with 26 groups registering for the automatic annotation task. In the end, 12 groups participated in the annotation task, submitting a total of 41 runs.

Automatic annotation of images in general, in particular of medical images, is a topic of great importance and relevance to the medical community. Currently the most relevant areas are:

1. *Automatic Parameter Setting for Image Analysis.* The variety of imaging modalities, appearances of different body regions, and the different diagnostic aims require medical image analysis to be specially adapted to the problem. For automatic chains of image processing and analysis, the processing modules must be parameterized accordingly. Thus, the need to classify images emerges to automatically select the necessary image processing steps.
2. *Consistency Checks for Meta Data.* Medical images are usually stored in the digital imaging and communication in medicine (DICOM) standard that also hosts various other meta data. However, a significant portion of this meta data are wrong, especially when it is generated automatically by the imaging device: Gld et al. (2002) reported that an approximate 15–20% of medical images that are recorded using DICOM-compliant modalities have incorrectly specified DICOM tags. In particular, coding of body region is frequently incorrect. Thus, another application of automatical classification of medical images is the validation and correction of their meta data.
3. *Generation of Text Queries for Retrieval.* In picture archiving and communication systems (PACS), information retrieval is based solely on alphanumeric attributes, i.e. text describing the patient, study, etc. With the increasing importance of images in daily medical routine, effective data management is required. By means of automatic image annotation, a textual description generated completely automatic from image content can be used to improve the query result. These tasks are strongly related to object recognition, and interestingly the methods from one discipline work surprisingly well for the respective other (Clough et al., 2005). Reflecting on these applications, the automatical image annotation task in the CLEF 2005 campaign aims at comparing and evaluating different approaches for automatically categorizing images.

2. The IRMA Database

In November 2005, the IRMA database¹² consisted of approximately 17,000 medical radiographs that have been

collected arbitrarily from daily routine at the Department of Diagnostic Radiology, RWTH Aachen University, Aachen, Germany.¹³

In order to establish a ground truth, the images were manually classified according to a mono-hierarchical, multi-axial coding scheme. More specifically, four axes are used to describe the technique (modality), orientation, body part, and biosystem (Lehmann et al., 2003). Each of these axes allows for specification in three or four levels of detail, and manual reference annotation was performed by skilled radiologists. This annotation process was partly computer-assisted by offering a pre-selection of the most likely annotations.

To ease participation, in the first year of automatic image annotation in ImageCLEF the complete IRMA code was not used. Instead, images were grouped (according to their annotation) at a coarser level of detail, forming 57 classes. An example image from each of these 57 classes is depicted in Fig. 1. All images were provided as PNG files, scaled to fit into a 512×512 pixel bounding box (keeping aspect ratio) using 256 gray values.

A subset of 10,000 images was used for ImageCLEF 2005. From this, a set of 9,000 randomly selected images (and category information) was selected as training data and given to registered participants prior to the evaluation. The remaining 1,000 images were later published as test data without category information to prevent *training on the testing data*. Performance was computed on the 1,000 test images, and systems compared according to their ability to correctly annotate these images.

According to radiology routine, the classes are unevenly distributed. For instance, the largest class (frontal chest radiographs) has a 28.6% (2860 images) share of the complete dataset, the second largest class makes up 9.6% (959 images) of the collection, and there are several classes that form only between 0.1% and 0.2% (10 to 20 images) of the complete set. However, the dataset was designed such that each class consists of at least 10 images. Clearly, some of the classes are visually very similar, e.g. class 7 (plain radiography, coronal, radio carpal joint, musculoskeletal system) and class 8 (plain radiography, coronal, hand, musculoskeletal system) where the only difference is the shown body part. In these classes, the body parts depicted are very similar as can be seen from the example images for classes 7 and 8 in Fig. 1.

3. Results

While ImageCLEF 2004 already attracted strong participation from 18 research groups across the world, in 2005, a total of 36 groups registered for ImageCLEF, with 26 groups registering for the automatic annotation task. In the end, 12 groups submitted results in the annotation task, leading to a total of 41 runs. The group

with the highest number of submissions had seven runs; one group submitted just one run. Table 1 lists participating groups and a short description of the methods used. References are given for more detailed description of methods.

As a baseline, the priori probability classifier, i.e. choose always the class with the highest number of observations in the training data, leads to an error rate of 71.1%. This error rate means that 711 of the 1000 images to be classified are misclassified. A more reasonable baseline for optical character recognition (OCR) and medical radiographs was suggested by Keyzers et al. (2003). It is provided by a nearest neighbor classifier comparing 32×32 256 gray level thumbnails of the images using the Euclidean distance. On this task, the nearest neighbor Euclidean distance classifier achieves an error rate of 36.8%. The best and worst error rate in the evaluation is 12.6% and 73.3%, respectively (Table 1). A combination from various classifiers submitted by participants could not improve over the best submission.

Obviously, the classification accuracy depends strongly on the specific class under consideration. The average classification accuracy over all runs for the different classes ranges from 6.3% to 90.7% and there is a tendency that classes with fewer training images are more difficult. For example, images from class 2 (“*plain radiography, coronal, facial cranium, musculoskeletal system*”) were frequently misclassified as class 44 (“*plain radiography, other orientation, facial cranium, musculoskeletal system*”): an average of 46% of images from class 2 were classified as class 44.

Classes 7 (“*plain radiography, coronal, radio carpal joint, musculoskeletal system*”) and 8 (“*plain radiography, coronal, hand/forearm, musculoskeletal system*”) are frequently misclassified as class 6 (“*plain radiography, coronal, hand, musculoskeletal system*”), where again class 6 is much better represented in the training data. Furthermore, many classes (6, 13, 14, 27, 28, 34, 44, 51, 57) are often misclassified to be from class 12, which is by far the largest class in the training data. This strongly coincides with the fact that class 12 (“*plain radiography, coronal, chest, unspecified*”) is the class with the highest classification accuracy: on average 90.7% of the test images from class 12 were classified correctly. The three classes with the lowest classification accuracies, form together less than 1% of the training data. In Fig. 2, the average confusion matrix over all submissions is visualized. Here, darker fields denote higher values. As the main diagonal of the matrix has much higher values than the other fields, the classifiers perform well on average

Three criteria are used to analyze the methods:

- *Image representation.* Several methods directly use the pixel values of the images and account for possible

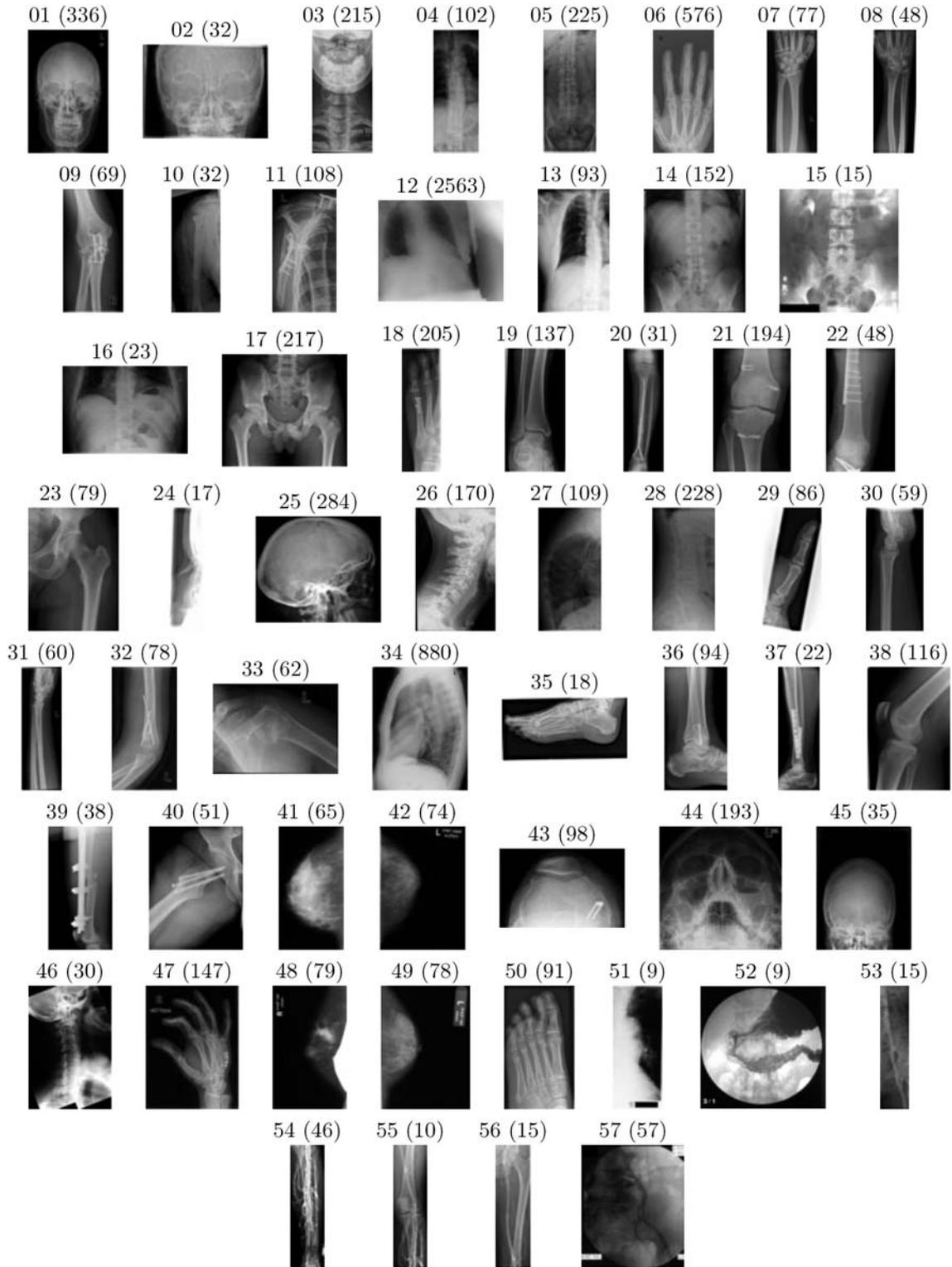


Figure 1. One example from each of the 57 classes (with the number of training examples from this class).

deformations in the images (i.e. ranks 1, 2, and 5). The methods coming from the object recognition field follow the currently widely adopted assumption that objects in images consist of parts that can be

modelled independently. Thus, these methods use local features extracted around interest points (i.e. ranks 3, 4, and 5). Other methods use quantization to different numbers of gray levels in combination with

Table 1. Resulting error rates for the submitted runs.

Rank	Group	Method	Ref.	ER[%]
1	RWTH Aachen U, CS Dep., DE	IDM	Keysers et al. (2004)	12.6
2	RWTH Aachen U, Med. Inf., DE	IDM & Texture feature	Lehmann et al. (2005)	13.3
3	RWTH Aachen U, CS Dep., DE	image patches & discriminative training	Deselaers et al. (2005)	13.9
4	U Liège, BE	image patches & boosting	Marée et al. (2005)	14.1
5	RWTH Aachen U, Med. Inf., DE	IDM & Texture feature	Lehmann et al. (2005)	14.6
6	U Liège, BE	image patches & decision trees	Marée et al. (2005)	14.7
7	U & Hospital Geneva, CH	MedGIFT	Müller et al. (2006)	20.6
8	Infocom, Singapore, SG	SVM & various image features	Qiu et al. (2005)	20.6
9	U & Hospital Geneva, CH	MedGIFT	Müller et al. (2006)	20.9
10	Infocom, Singapore, SG	SVM & various image features	Qiu et al. (2005)	20.9
11	Infocom, Singapore, SG	SVM & various image features	Qiu et al. (2005)	21.0
12	U & Hospital Geneva, CH	MedGIFT	Müller et al. (2006)	21.2
13	U & Hospital Geneva, CH	MedGIFT	Müller et al. (2006)	21.3
14	Miracle from UPM Madrid, ES	GIFT & majority voting	Villena-Román et al. (2005)	21.4
15	U & Hospital Geneva, CH	MedGIFT	Müller et al. (2006)	21.7
16	Infocom, Singapore, SG	SVM & various image features	Qiu et al. (2005)	21.7
17	National Taiwan U, TW	block features & nearest neighbor	Chang et al. (2005)	21.7
18	National Taiwan U, TW	block features & top 2 classifier	Chang et al. (2005)	21.7
19	U & Hospital Geneva, TW	MedGIFT	Müller et al. (2006)	21.8
20	U & Hospital Geneva, TW	MedGIFT	Müller et al. (2006)	22.1
21	Miracle from UPM Madrid, ES	GIFT & majority voting	Villena-Román et al. (2005)	22.3
22	National Taiwan U, TW	block features & nearest prototype	Chang et al. (2005)	22.5
23	DBLAB from NCTU, TW	SVM, various img. feat.	Cheng et al. (2005)	24.7
24	DBLAB from NCTU, TW	SVM, various img. feat.	Cheng et al. (2005)	24.9
25	DBLAB from NCTU, TW	SVM, various img. feat.	Cheng et al. (2005)	28.5
26	DBLAB from NCTU, TW	SVM, various img. feat.	Cheng et al. (2005)	31.8
27	DBLAB from NCTU, TW	SVM, various img. feat.	Cheng et al. (2005)	33.8
28	CEA, FR	k -NN classifier & image projection	Besançon and Millet (2005)	36.9
29	Mt. Holyoke College, MA, USA	Gabor Energy	Petkova and Ballesteros (2005)	37.8
30	Mt. Holyoke College, MA, USA	Gabor Energy	Petkova and Ballesteros (2005)	40.3
31	CEA, FR	k -NN & local edge patterns	Besançon and Millet (2005)	42.5
32	CINDI from Concordia U, CA	SVM, various image feat.	Rahman et al. (2005)	43.3
33	CEA, FR	k -NN & quantified colors	Besançon and Millet (2005)	46.0
34	U Montreal, CA	feature combination		55.7
35	U Montreal, CA	texture coarseness		60.3
36	DBLAB from NCTU, TW	SVM, various img. feat.	Cheng et al. (2005)	61.5
37	U Montreal, CA	contour image features		66.6
38	U Montreal, CA	shape image features		67.0
39	U Montreal, CA	centered contours		67.3
40	U Montreal, CA	Fourier shape feat.		67.4
41	U Montreal, CA	directionality		73.3
	Euclidean Distance, 32×32 images, 1-Nearest-Neighbor			36.8
	apriori probability classifier			71.1

Abbreviations for groups: *U*: University, *CS*: computer science, *UPM*: Universidad Politecnica de Madrid, *DBLAB*: database lab, *NCTU*: National Chiao Tung University, *NTU*: National Taiwan University, *CEA*: Commissariat à l'énergie atomique. Abbreviations for methods: *IDM*: image distortion model, *SVM*: support vector machines, *MedGIFT*: medical GNU image finding tool, *GIFT*: GNU image finding tool, *k-NN*: k nearest neighbor.

Gabor filters (ranks 7, 9, 12, 13, 15, 19, and 20) from the medGIFT¹⁴ image retrieval system. As the images do not contain any color information, texture features like the Tamura texture features Tamura et al. (1978), the MPEG-7 visual descriptors (Eidenberger, 2003), or Gabor features (Gabor, 1946) play an important role for this task and were used by several groups (e.g. ranks 2, 5, 7, 8, 9, 10, ...).

- *Classification Method*. Many of the submitted results were created using k -nearest neighbor classification

with k between 1 and 20 (ranks 1, 2, 5, 7, 17, ...). Several methods use the GIFT retrieval metric for the determination of nearest neighbors (ranks 7, 9, 12, 13, 15, 19, 20, ...). Variations like nearest prototype classification and majority voting were also applied. One method uses a maximum entropy classifier (rank 3), one method uses boosting and decision trees (ranks 4 and 6), and some groups use support vector machines (ranks 8, 10, 11, 16, 23, 24, 25, 26, 27, 32, and 36). For support vector machines, the variety

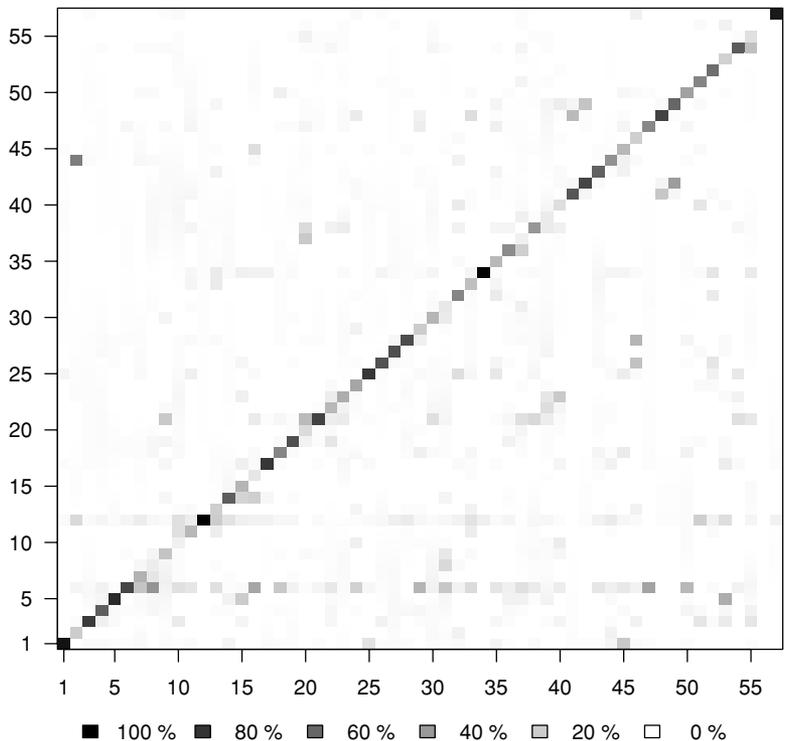


Figure 2. Average confusion matrix over all submitted runs. Dark fields denote high values.

in performance is high and probably depends on the image representations and kernels used.

- *Learning Technique.* Explicit learning and training methods were used only by the groups using support vector machines, boosting and decision trees, or the maximum entropy classifier. The other groups used the parameter-free nearest neighbor classifier. Nonetheless, the medGIFT group is planning to use a

training method in upcoming annotation tasks (Müller et al., 2004).

In summary, it can be seen that the best results are obtained using the pixel values of the images directly: either by using sparsely sampled image patches or by using the complete image unchanged. For classification, discriminative approaches seem to perform very well and nearest

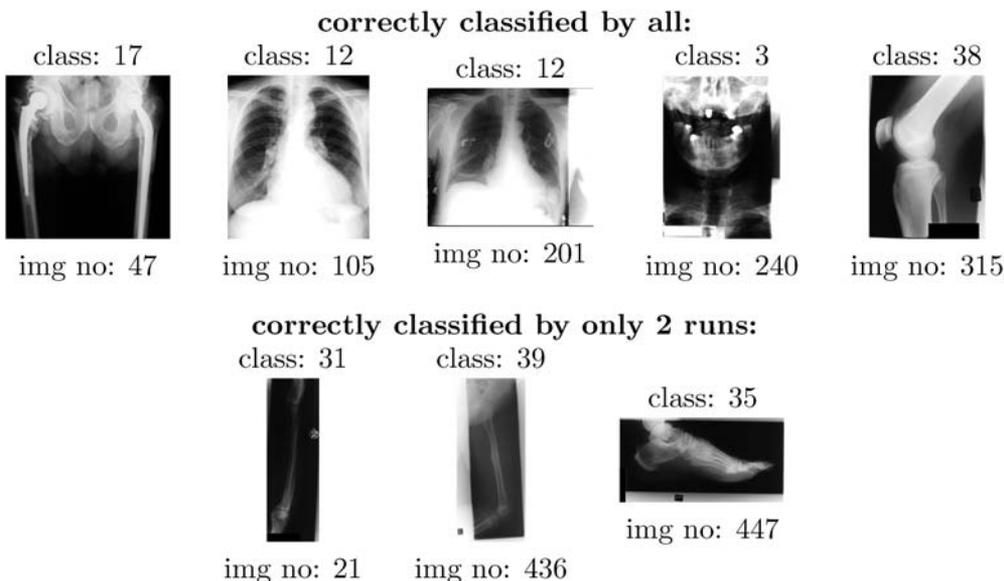


Figure 3. The images that were classified correctly by all systems and the images that were misclassified most often.

neighbor classification is also well suited if a suitable distance function can be defined.

In Fig. 3 some test images that were correctly classified by all submissions and some test images that were misclassified most frequently are depicted together with their classes.

Discussion. Due to the high participation and the good results that were achieved by several methods, the automatic annotation task in ImageCLEF 2005 appears to have been a great success. The task can be considered to be realistic, as the images were taken randomly from clinical routine and the problem of correcting annotations of primarily digital images and annotation of secondary digital images is a daily problem. In summary, the database is a valuable resource for testing and creating automatic image annotation systems, and the high participation in ImageCLEF has shown the need for such an evaluation.

Although most of the participating methods come from a CBIR context (e.g. the methods ranked 1, 2, 5, 7, 9, 12, 13, 14, ...) it can be seen that those methods that come from the image classification and recognition domain field (e.g. the methods ranked 3, 4, 6, 8, 10, 11) can achieve excellent results for the task of automatic annotation of medical images. The success of the two groups from RWTH Aachen University might partly be due to their working with similar data for some time before.

Methods using the pixel values directly and deformation models outperform most other methods for the given task. Methods from object recognition, assuming that objects in images can be modelled as a set of parts, also perform very well although they were not tuned with respect to this task. It can also be seen that image retrieval methods perform well for this task, especially if domain knowledge from medicine can be incorporated (ranks 2, 5, 7, ...).

Interestingly, the classifier used is not of such great importance, because the classifiers that were applied are spread over the whole range of submissions. The methods from object recognition have the advantage that training of these methods is a well-investigated area and usually discriminative methods are applied. In contrast, in the CBIR domain the training of parameters is still uncommon.

4. Summary and conclusion

We presented the outcomes of the medical automatic image annotation task of ImageCLEF 2005 and briefly described the methods of the participating groups. The data of the evaluation is available free of charge to encourage comparison of new approaches to the outcomes of the challenge.

Outlook. With ImageCLEFmed 2005, a valuable resource for benchmarking of automatic annotation algorithms has been created. In 2006, this task was extended and continued with a similar number of participants. 10,000 training images from 117 classes were provided and a new set of 1,000 test images. Furthermore, a non-medical automatic annotation task was established in cooperation with the MUSCLE¹⁵ network of excellence.¹⁶

For 2007, it is planned to create a *hierarchical* classification task. That is, the training images are published with their entire IRMA code and a new set of test images has to be classified. Instead of absolute decisions for a class, the classifier will be able to decide on its own to what level of detail the classification is done on which of the annotation axes.

Acknowledgment

This work was partially funded by the Deutsche Forschungsgemeinschaft (DFG) under grants NE-572/6 and Le 1108/4 and by the Swiss National Science Foundation (NSF) grant ITR-0325160 and grant 205321-109304. Furthermore we would like to thank the CLEF campaign for allowing ImageCLEF to be part of it.

Notes

1. <http://www.nist.gov/speech/tests/mt/>.
2. <http://tc-star.org>.
3. <http://www.is.cs.cmu.edu/iwslt2005/evaluation.html>.
4. <http://trec.nist.gov>.
5. <http://www.benchathlon.net/>.
6. <http://www.imageval.org/>.
7. <http://www-nlpir.nist.gov/projects/trecvid/>.
8. <http://trec.nist.gov/>.
9. <http://www.pascal-network.org/>.
10. <http://www.clef-campaign.org>.
11. <http://ir.shef.ac.uk/imageclef/>.
12. <http://irma-project.org>.
13. <http://www.rad.rwth-aachen.de>.
14. <http://www.sim.hcuge.ch/medgift/>.
15. Multimedia Understanding through Semantics, Computation and Learning.
16. <http://www.muscle-noe.org/>.

References

- Ballesteros, L. and Petkova, D. 2005. A clustered retrieval approach for categorizing and annotating images. In *Working Notes of the CLEF Workshop 2005*, Vienna, Austria, LNCS 4022, pp. 662–672.
- Besaçon, R. and Millet, C. 2005. Data fusion of retrieval results from different media: Experiments at ImageCLEF 2005. In *Working Notes of the CLEF Workshop 2005*, Vienna, Austria, LNCS 4022, pp. 622–631.
- Chang, Y.C., Lin, W.C., and Chen, H.H. 2005. A corpus-based relevance feedback approach to cross-language image retrieval. In *Working*

- Notes of the CLEF Workshop 2005*, Vienna, Austria, LNCS 4022, pp. 592–601.
- Chen, D.R., Chang, R.F., and Huang, Y.L. 1999. Computer-aided diagnosis applied to use of solid breast nodules by using neural networks. *Radiology*, 213:407–412.
- Cheng, P.C., Chien, B.C., Ke, H.R., and Yang, W.P. 2005. Combining textual and visual features for cross-language medical image retrieval. In *Working Notes of the CLEF Workshop 2005*, Vienna, Austria, LNCS 4022, pp. 712–723.
- Clough, P., Müller, H., Deselaers, T., Grubinger, M., Lehmann, T.M., Jensen, J., et al. 2005. The CLEF 2005 cross-language image retrieval track. In *Workshop of the Cross-Language Evaluation Forum (CLEF 2005)*, Vienna, Austria, LNCS 4022, pp. 535–557.
- Deselaers, T., Keysers, D., and Ney, H. 2005. Discriminative training for object recognition using image patches. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 05)*, San Diego, CA, Vol. 2, pp. 157–162.
- Eidenberger, H. 2003. How good are the visual MPEG-7 features? In *Proceedings SPIE Visual Communications and Image Processing Conference 2003*; Lugano, Italy, Vol. 5150, pp. 476–488.
- Everingham, M., Zisserman, A., Williams, C.K.I., van Gool, L., Allan, M., et al. 2006. The 2005 PASCAL visual object classes challenge. In *Selected Proceedings of the First PASCAL Challenges Workshop. Lecture Notes in Artificial Intelligence*, Southampton, UK, Springer, LNAI 3944, pp. 117–176.
- Gabor, D. 1946. Theory of communication. *J. IEE*, 93:429–457.
- Güld, M.O., Kohnen, M., Keysers, D., Schubert, H., Bredno, J., and Lehmann, T.M. 2002. Quality of DICOM header information for image categorization. In *SPIE 2002*, Vol. 4685, pp. 280–287.
- Keysers, D., Dahmen, J., Ney, H., Wein, B., and Lehmann, T.M. 2003. Statistical framework for model-based image retrieval in medical applications. *J. Electron. Imaging*, 12(1):59–68.
- Keysers, D., Gollan, C., and Ney, H. 2004. Classification of medical images using non-linear distortion models. In *Proc. BVM 2004, Bildverarbeitung für die Medizin*, Berlin, Germany, pp. 366–370.
- Lehmann, T.M., Güld, M.O., Deselaers, T., Keysers, D., Schubert, H., Spitzer, K., Ney, H., and Wein, B.B. 2005. Automatic categorization of medical images for content-based retrieval and data mining. *Comput. Med. Imaging and Graphics*, 29:143–155.
- Lehmann, T.M., Schubert, H., Keysers, D., Kohnen, M., and Wein, B.B. 2003. The IRMA code for unique classification of medical images. In *Proc. SPIE 2003*, Vol. 5033, pp. 440–451.
- Marée, R., Geurts, P., Piater, J., and Wehenkel, L. 2005. Biomedical image classification with random subwindows and decision trees. In *Proc. ICCV workshop on Computer Vision for Biomedical Image Applications (CVIBA 2005). Lecture Notes in Computer Science 3765*, pp. 220–229.
- Martínez-Fernández, J.L., Villena-Roman, J., García-Serrano, A.M., and Gonzales-Cristobal, J.C. 2005. Combining textual and visual features for image retrieval. In *Working Notes of the CLEF Workshop 2005*, Vienna, Austria, LNCS 4022, pp. 680–691.
- Müller, H., Geissbühler, A., Marty, J., Lovis, C., and Ruch, P. 2006. The use of medGIFT and easyIR for ImageCLEF 2005. In *Proceedings of the Cross Language Evaluation Forum 2005. Lecture Notes in Computer Science*, Vienna, Austria, LNCS 4022, pp. 724–731.
- Müller, H., Michoux, N., Bandon, D., and Geissbühler, A. 2004. A review of content-based image retrieval systems in medical applications—clinical benefits and future directions. *International Journal of Medical Informatics*, 73:1–23.
- Müller, H., Squire, D.M., and Pun, T. 2004. Learning from user behaviour in image retrieval: applications of the market basket analysis. *Int. J. Comput. Vision*, 56(1–2) (Special Issue on Content-based Visual Information Retrieval):65–77.
- Rahman, M., Desai, B.C., and Bhattacharya, P. 2005. Supervised machine learning based medical image annotation and retrieval. In *Working Notes of the CLEF Workshop 2005*, Vienna, Austria, LNCS 4022, pp. 692–701.
- Tamura, H., Mori, S., and Yamawaki, T. 1978. Textural features corresponding to visual perception. *IEEE Trans. Syst., Man, Cybernetics*, 8(6):460–472.
- Tourassi, G.D. 1999. Journey toward computer-aided diagnosis: role of image texture analysis. *Radiology*, 213:317–320.
- Xiong, W., Qiu, B., Tian, Q., Xu, C., Ong, S.H., and Foong, K. 2005. Combining visual features for medical image retrieval and annotation. In *Working Notes of the CLEF Workshop 2005*, Vienna, Austria, LNCS 4022, pp. 632–641.
- Yamamoto, S., Jiang, H., Matsumoto, M., Tateno, Y., Iinuma, T., and Matsumoto, T. 1996. Image processing for computer-aided diagnosis of lung cancer by CT. In *3rd IEEE Workshop on Applications of Computer Vision*, Sarasota, FL, USA, pp. 236–241.