Medical Image Annotation in ImageCLEF 2008

Thomas Deselaers[1] and Thomas M. Deserno[2]

[1] RWTH Aachen University, Computer Science Department, Aachen, Germany
deselaers@cs.rwth-aachen.de
[2] RWTH Aachen University, Dept. of Medical Informatics, Aachen, Germany
deserno@ieee.org

Abstract. The ImageCLEF 2008 medical image annotation task is designed to assess the quality of content-based image retrieval and image classification by means of global signatures. In contrast to the previous years, the 2008 task was designed such that the hierarchy of reference IRMA code classifications is essential for good performance. In total, 12,076 images were used, and 24 runs of 6 groups were submitted. Multi-class classification schemes for support vector machines outperformed the other methods. A scoring scheme was defined to penalise wrong classification in early code positions over those in later branches of the code hierarchy, and to penalise false category association over the assignment of a "not known" code. The obtained scores rage from 74.92 over 182.77 to 313.01 for best, baseline and worst results, respectively.

1 Introduction

From the first introduction of the medical image annotation task in ImageCLEF to now this task evolved form a simple classification task with only about 60 classes [3] to a task with nearly 120 classes [6] and further to a task where a complex class hierarchy of potentially several thousand classes had to be considered [4].

In 2005, the aim of the medical image annotation task was defined as exploring and promoting the use of automatic annotation techniques to for extracting semantic information from little-annotated medical images. Therefore a new database of 10,000 images from 57 classes was created. This database was extended each year by adding at least 1,000 images. Furthermore the difficulty of the classification was increased by first increasing the number of classes and later including a complex hierarchical class structure: the Image Retrieval in Medical Applications (IRMA) code [5]. However, even the 2007 task could be solved using flat classification hierarchies since large parts of the hierarchy were unused and the effective number of classes was only slightly higher than in 2006.

With the 2008 task, we have achieved the goals that were set out initially: an image annotation task which requires the explicit use of the class-hierarchy in order to achieve good results and a wide variety of different methods has been systematically evaluated by the participating groups.

Other tracks in ImageCLEF 2008 were the photo retrieval task [1], the medical retrieval task [7], the Wikipedia multimedia retrieval task [8], and the visual concept detection task [2].

2 Materials and Methods

The aim of the 2008 medical image annotation task was to promote the use of hierarchical classification techniques and foster the use of the prior knowledge encoded into the hierarchy of classes. Thus, the task was similar to the task of 2007 in that the classes were based on the IRMA code [5]. The main difference this year was that the prior distribution of the classes in the test data differed strongly from the prior distribution of the training data and that thus in particular classes which were badly represented in the training data were present in the test data to encourage the use of the hierarchy and the placement of wild card operators.

2.1 Database and Task Description

The training data of this year consisted of 12,076 images (10,000 training images from last year + 1,000 development images from last year + 1,000 test images from last year + 76 new images) and the test data consisted of 1,000 new images. In total 196 unique codes were present in the training images and 187 of these were present in the test images. The most frequent class in the training data consisted of more than 2,300 images, but the test data had only one example from this class. In Figure 1, the frequency of classes in the training and in the test data is shown. It can be seen that the classes in the test data were nearly uniformly distributed, but, in the training data, some classes were far more frequent than others.

Each of the radiographs is annotated with its complete IRMA code (see Sec. 2.2). In total, 196 different IRMA codes occurred in the database. Example images from the database together with textual labels and their complete code are given in Figure 2 and 3.

2.2 IRMA Code

Existing medical terminologies such as the MeSH thesaurus are poly-hierarchical, i.e., a code entity can be reached over several paths. However, in the field of content-based image retrieval, we frequently find class-subclass relations. The mono-hierarchical multi-axial IRMA code strictly relies on such part-of hierarchies and, therefore, avoids ambiguities in textual classification [5]. In particular, the IRMA code is composed from four axes having three to four positions, each in $\{0, \ldots 9, a, \ldots z\}$, where "0" denotes "not further specified". More precisely,

- the technical code (T) describes the imaging modality;
- the directional code (D) models body orientations;

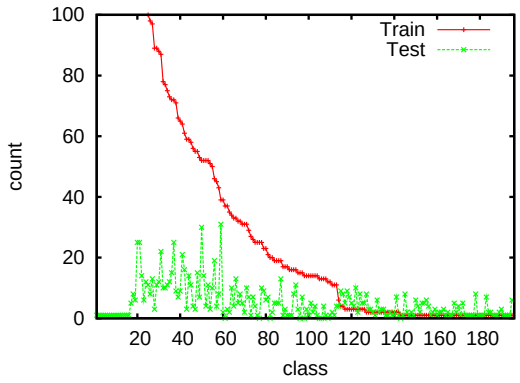


Fig. 1: Frequency of images in the training and test data.

– the anatomical code (A) refers to the body region examined; and
– the biological code (B) describes the biological system examined.

This results in a string of 13 characters (IRMA: TTTT – DDD – AAA – BBB). A small exemplary excerpt from the anatomy axis of the IRMA code is given in Table 1.

The IRMA code can be easily extended by introducing characters in a certain code position, e.g., if new imaging modalities are introduced. Based on the hierarchy, the more code position differ from "0", the more detailed is the description.

2.3 Hierarchical Classification

Let an image be coded by the above 4 *independent* axes, such that we can consider the axes independently and just sum up the errors for each axis independently:

- let $l_1^I = l_1, l_2, \ldots, l_i, \ldots, l_I$ be the *correct* code (for one axis) of an image;
- let $\hat{l_1^I} = \hat{l_1}, \hat{l_2}, \ldots, \hat{l_i}, \ldots, \hat{l_I}$ be the *classified* code (for one axis) of an image;

where l_i is specified precisely for each position, and in $\hat{l_i}$ it is allowed to say "*don't know*", which is encoded by $*$. Note that I (the depth of the tree to which the classification is specified) may be different for different images.

Given an incorrect classification at position $\hat{l_i}$ we consider all succeeding decisions to be wrong and given a not specified position, we consider all succeeding decisions to be not specified. Furthermore, we do not count any error if the correct code is unspecified and the predicted code is a wildcard. In that case, we do consider all remaining positions to be not specified.

Since we want to penalise wrong decisions that are easy (fewer possible choices at that node) over wrong decisions that are difficult (many possible

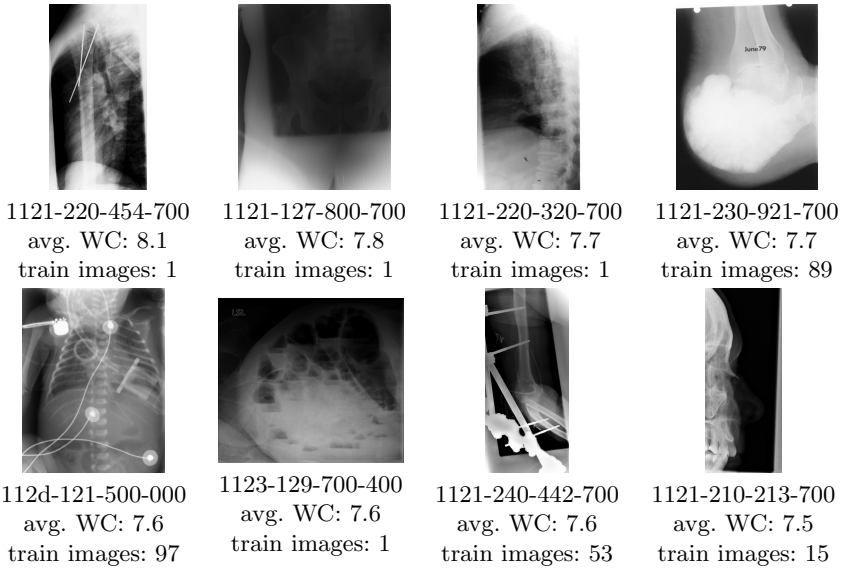


Fig. 2: The test images from the database where most wildcards were used with their full IRMA code and the average number of wildcards over all runs.

choices at that node), a decision at position l_i is considered to be correct by chance with a probability of $\frac{1}{b_i}$, if b_i is the number of possible labels for position i. This assumes equal priors for each class at each position.

Furthermore, we want to penalise wrong decisions at an early stage in the code (higher up in the hierarchy) over wrong decisions at a later stage in the code (lower down on the hierarchy) (i.e. l_i is more important than l_{i+1}).

Putting this together yields:

$$\sum_{i=1}^{I} \underbrace{\frac{1}{b_i}}_{(a)} \underbrace{\frac{1}{i}}_{(b)} \underbrace{\delta(l_i, \hat{l}_i)}_{(c)} \tag{1}$$

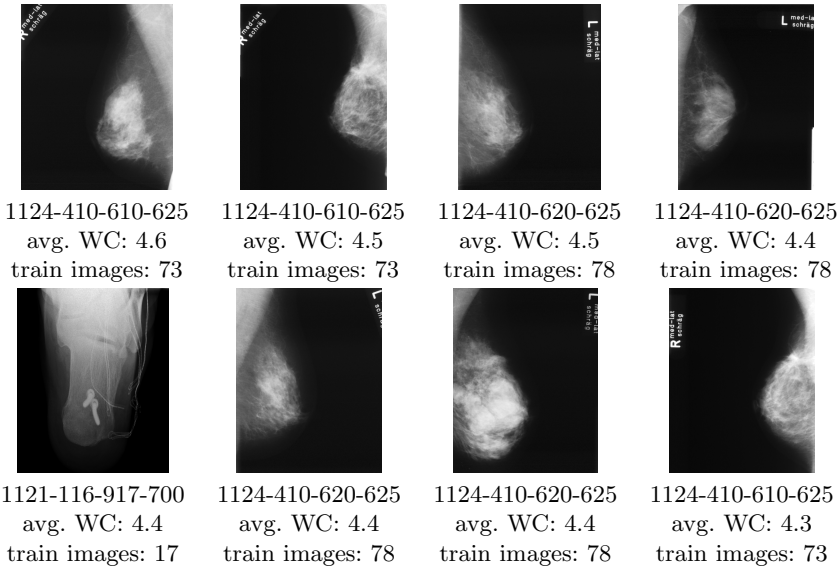
with

$$\delta(l_i, \hat{l}_i) = \begin{cases} 0 & \text{if } l_j = \hat{l}_j \quad \forall j \leq i \\ 0.5 & \text{if } l_j = * \quad \exists j \leq i \\ 1 & \text{if } l_j \neq \hat{l}_j \quad \exists j \leq i \end{cases}$$

where the parts of the equation account for

(a) difficulty of the decision at position i (branching factor);
(b) the level in the hierarchy (position in the string); and
(c) the correct/not specified/wrong labelling, respectively.

In addition, for each axis, the maximal possible error is calculated and the errors are normed such that a completely wrong decision (i.e. all positions for



1124-410-610-625 avg. WC: 4.6 train images: 73
1124-410-610-625 avg. WC: 4.5 train images: 73
1124-410-620-625 avg. WC: 4.5 train images: 78
1124-410-620-625 avg. WC: 4.4 train images: 78
1121-116-917-700 avg. WC: 4.4 train images: 17
1124-410-620-625 avg. WC: 4.4 train images: 78
1124-410-620-625 avg. WC: 4.4 train images: 78
1124-410-610-625 avg. WC: 4.3 train images: 73

Fig. 3: The test images from the database where fewest wildcards were used with their full IRMA code and the average number of wildcards over all runs.

that axis are wrong) gets an error count of 0.25, and a completely correctly predicted axis has an error of 0. Thus, an image where all positions in all axes are wrong has an error count of 1, and an image where all positions in all axes are correct has an error count of 0. An example of this scheme is given in Table 2.

3 Results from the Evaluation

In 2008, 6 groups participated in the medical annotation task submitting 24 runs in total. In the following, we briefly describe the methods applied by the participating groups.

FEIT. The Faculty of Electrical Engineering and Information Technologies from the University of Skopje in Macedonia submitted two runs using global and local image descriptors, which are classified using bagging and random forests.

medGIFT. The medical Gnu Image Finding Tool (medGIFT) group from University Hospitals of Geneva in Switzerland submitted four runs using different descriptors and voting schemes in the medGIFT image retrieval system.

Miracle. The Miracle group from Daedalus University in Spain submitted four runs using different global and local image descriptors in a nearest neighbour classifier.

TAU-BIOMED. The Biomedical Image Processing Lab from Tel Aviv University in Israel submitted four runs using a bag-of-visual words approach with dense sampling and support vector machines for classification.

Table 1: Examples from the IRMA code

AAA code	textual description
000	not further specified
...	
400	upper extremity (arm)
410	upper extremity (arm); hand
411	upper extremity (arm); hand; finger
412	upper extremity (arm); hand; middle hand
413	upper extremity (arm); hand; carpal bones
420	upper extremity (arm); radio carpal join ...
430	upper extremity (arm); forearm
431	upper extremity (arm); forearm; distal forearm
432	upper extremity (arm); forearm; proximal forearm
440	upper extremity (arm); ellbow
...	

IDIAP. The "Institut Dalle Molle d'Intelligence Artificielle Perceptive" (IDIAP) Research Institute from Switzerland submitted nine runs using different multi-class classification schemes for support vector machines and different image descriptors.

RWTH-MI. The Image Retrieval in Medical Applications (IRMA) group at the Department of Medical Informatics, RWTH Aachen University in Aachen, Germany, provides a baseline-run that was computed using Tamura Texture Measures and the Image Distortion Model. Since 2004, the parameterisation remains unchanged, and, therefore, the hierarchy was disregarded.

The results from the evaluation are given in Table 3 sorted by error score. It can be seen that the classification accuracy varies strongly from 74.9 to 313 error points according to the above described error measurement. Also, the number of wildcards used varies very strongly between 0 in the model free approach from the IRMA group up to about 7,000, which means that almost seven wildcards per image were used on the average, i.e. more than half of the positions for the images are undefined.

In general, it an be seen that the discriminative models using local descriptors from the IDIAP group outperform the other approaches.

In Figures 2 and 3, some example test images are given along with their full IRMA code. The number of wildcards used by the submitted runs on average and the number of training images from this particular class. The top and the bottom parts of the figure show the images where, on the average, the most and the fewest wildcards were used, respectively. It can be observed that for classes with bad support in the training data far more wildcards were used.

Table 2: Example for different errors in the hierarchical classification scheme. Assuming the code `318a` is correct.

predicted code	error score
318a	0.0
318*	0.0
3187	0.0
31*a	0.1
31**	0.1
3177	0.2
3***	0.3
32**	0.7
1000	1.0

4 Discussion and Conclusion

We have presented the ImageCLEF 2008 medical image annotation task. In contrast to previous years, the distribution of training and test images was chosen such that using the hierarchy of the IRMA code was necessary to obtain good results. For classes with very few training images, the submitted runs employed up to more than eight wildcards out of thirteen code positions per image to express their uncertainty about the classifications. Multi-class classification schemes for support vector machines, as used by the IDIAP Research Institute of Switzerland, outperformed the other methods. The obtained scores rage from 74.92 over 182.77 to 313.01 for best, baseline and worst, respectively.

In total the goals initially setup for the medical image annotation task were achieved: techniques for the annotation of medical images were systematically evaluated on a series of tasks of gradually increasing difficulty and still the results of the best system was improved over the years. The medical image annotation will not be continued in ImageCLEF in its current form but hopefully new and challenging tasks will be proposed and offered.

References

1. Thomas Arni, Paul Clough, Mark Sanderson, and Michael Grubinger. Overview of the ImageCLEFphoto 2008 photographic retrieval task. In *Proceedings of the CLEF Workshop 2008*, Lecture Notes in Computer Science, Aarhus, Denmark, Sep 2008 (printed in 2009).
2. Thomas Deselaers and Allan Hanbury. The visual concept detection task in ImageCLEF 2008. In *Proceedings of the CLEF Workshop 2008*, Lecture Notes in Computer Science, Aarhus, Denmark, Sep 2008 (printed in 2009).
3. Thomas Deselaers, Henning Müller, Paul Clogh, Hermann Ney, and Thomas M Lehmann. The CLEF 2005 automatic medical image annotation task. *International Journal of Computer Vision*, 74(1):51–58, August 2007.

Table 3: Results from the medical image annotation task.

group	run	error score	wildcards
idiap	LOW_MULT_2MARG	74.92	4148
idiap	LOW_MULT	83.45	3154
idiap	LOW_2MARG	83.79	4353
idiap	MCK_MULT_2MARG	85.91	4655
idiap	LOW_lbp_siftnew	93.20	3157
idiap	SIFTnew	100.27	3144
TAU	BIOMED-svm_full	105.75	1000
TAU	BIOMED-svm_prob	105.86	4868
TAU	BIOMED-svm_vote	109.37	1000
TAU	BIOMED-svm_small	117.17	1000
idiap	LBP	128.58	3173
rwth_mi	baseline	182.77	0
MIRACLE	MIRACLE-3I-0F	187.90	4426
MIRACLE	MIRACLE-2I-0F	190.38	3194
MIRACLE	MIRACLE-2I-2F	190.38	3194
MIRACLE	MIRACLE-3I-2F	194.26	3871
GE	GIFT0.9_0.5_vcad_5	210.93	2146
GE	GIFT0.9_0.5_vca_5	217.34	2466
idiap	MCK_pix_sift_2MARG	227.82	6994
GE	GIFT0.9_akNN_2	241.11	1000
GE	GIFT0.9_kNN_2	251.97	1000
FEIT	1	286.48	1117
FEIT	2	290.50	1024
idiap	MCK_pix_sift	313.01	3420

4. Thomas Deselaers, Henning Müller, and Thomas M. Deserno. Automatic medical image annotation in ImageCLEF 2007: Overview, results, and discussion. *Pattern Recognition Letters*, page in press, 2008.
5. Thomas M. Lehmann, Henning Schubert, Daniel Keysers, M Kohnen, and Berthold B Wein. The IRMA code for unique classification of medical images. In *Proceedings SPIE*, volume 5033, pages 440–451, 2003.
6. Henning Müller, Thomas Deselaers, Thomas M. Lehmann, Paul Clough, and William Hersh. Overview of the ImageCLEFmed 2006 medical retrieval and annotation tasks. In *Evaluation of Multilingual and Multi-modal Information Retrieval – Seventh Workshop of the Cross-Language Evaluation Forum, CLEF 2006*, volume 4730, pages 595–608, Alicante, Spain, 2007.
7. Henning Müller, Jayashree Kalpathy-Cramer, Charles E. Kahn Jr., William Hatt, Steven Bedrick, and William Hersh. Overview of the ImageCLEFmed 2008 medical image retrieval task. In *Proceedings of the CLEF Workshop 2008*, Lecture Notes in Computer Science, Aarhus, Denmark, Sep 2008 (printed in 2009).
8. Theodora Tsikrika and Jana Kludas. Overview of the wikipediaMM task at ImageCLEF 2008. In *Proceedings of the CLEF Workshop 2008*, Lecture Notes in Computer Science, Aarhus, Denmark, Sep 2008 (printed in 2009).