# Medical Image Annotation in ImageCLEF 2008

Thomas Deselaers[1] and Thomas M. Deserno[2]
[1]RWTH Aachen University, Computer Science Department, Aachen, Germany
[2]RWTH Aachen University, Dept. of Medical Informatics, Aachen, Germany
deselaers@cs.rwth-aachen.de, 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 total, 12,076 images were used. In contrast to previous years, the task was designed such that the hierarchy of reference IRMA code classifications is essential for good performance. 24 runs of 6 groups were submitted. Multi-class classification schemes for support vector machines outperformed the other methods. The obtained scores rage from 74.92 over 182.77 to 313.01 for best, baseline and worst results, respectively.

## Keywords

ImageCLEF, medical image annotation, image classification

## 1 Introduction

Over the last three years, automatic medical image annotation evolved form a simple classification task with only about 60 classes [1] to a task with nearly 120 classes [4] and further to a task where a complex class hierarchy of potentially several thousand classes had to be considered [2]. 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.

The aim of this years medical image annotation task is to promote the use of hierarchical classification hierarchies and foster the use of the prior knowledge encoded into the hierarchy of classes.

The task of this year is similar to last year in that the classes are again based on the IRMA code [3]. The main difference this year is that the prior distribution of the classes in the test data is strongly different from the prior distribution of the training data and that thus in particular classes which are badly represented in the training data are present in the test data to encourage the use of the hierarchy and the placement of wild card operators.

## 2 Database and Task Description

The training data of this year consists 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 consists of 1,000 new images. In total 196 unique codes are present in the training images and 187 of these are present in the test images. The most frequent class in the training data consists of more than 2,300 images, but the test data has 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 are nearly uniformly distributed but in the training data some classes are far more frequent than others.

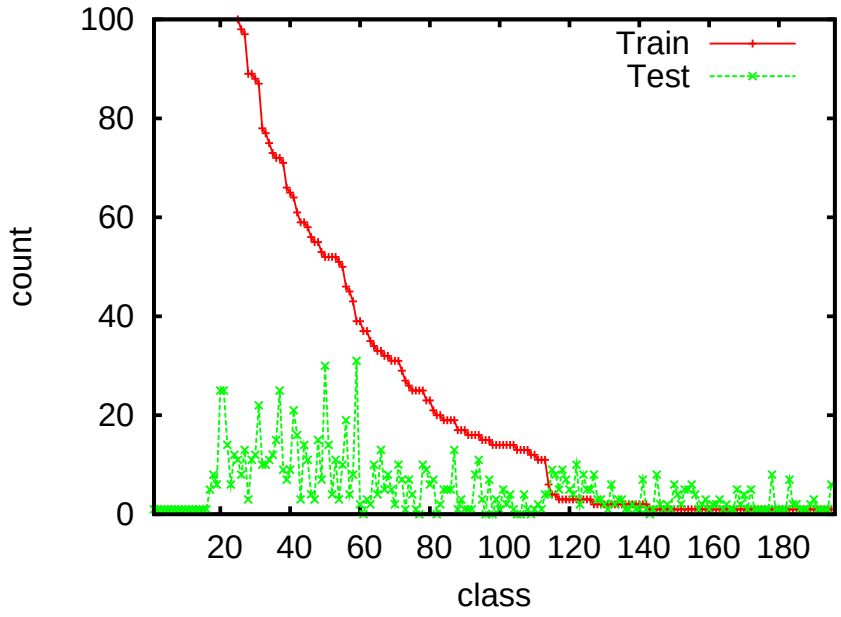



Figure 1: Frequency of images in the training and test data.

Each of the radiographs is annotated with its complete IRMA code (see Sec. 2.1). In total, 196 different IRMA codes occur in the database. Example images from the database together with textual labels and their complete code are given in Figure 2.

## 2.1 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 [3]. 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;
- 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.

## 2.2 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;

Table 1: Examples from the IRMA code

| 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 |
| ... | |

- 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 every 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 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.

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

| | |
|---|---|
| `318a` | 0.0 |
| `318*` | 0.0244653860094 |
| `3187` | 0.0489307720188 |
| `31*a` | 0.0824574121058 |
| `31**` | 0.0824574121058 |
| `3177` | 0.164914824212 |
| `3***` | 0.34342152954 |
| `32**` | 0.686843059079 |
| `1000` | 1.0 |

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

# 3 Results from the Evaluation

In 2008, only 6 groups participated in the medical annotation task submitting 24 runs in total. In the following, we 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 2 runs using global and local image descriptors which are classified using bagging and random forests.

**medGIFT.** The medGIFT group from University Hospitals of Geneva in Switzerland submitted 4 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 Medical 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.

**IDIAP.** The IDIAP research institute from Switzerland submitted 9 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 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 parameterization remains unchanged, and, therefore, the hierarchy was disregarded.

The results from the evaluation are given in Table 3 sorted by error score. It an be seen that the classification accuracy varies strongly from 74.9 errors to 313 errors 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 to up to 7000, which means that almost 7 wildcards per image were used on the average, i.e. more than half of the positions for the images are undefined.

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 |

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

In Figure 2, 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-part and the bottom-part 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.

# 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 8 wildcards out of 13 code positions per image to express their uncertainty about certain 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.

# References

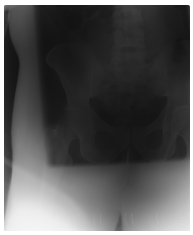

[1] Thomas Deselaers, Henning Müller, Paul Clogh, Hermann Ney, and Thomas M Lehmann. The clef 2005 automatic medical image annotation task. *International Journal of Computer Vision*, 74(1):51–58, August 2007.

[2] 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.

[3] 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.

[4] Henning Müller, Thomas Deselaers, Thomas M. Lehmann, Paul Clough, and William Hersh. Overview of the imageclefmed 2006 medical retrieval and annotation tasks. In C. Peters, P. Clough, F. Gey, J. Karlgren, B. Magnini, D.W. Oard, M. de Rijke, and M. Stempfhuber, editors, *Evaluation of Multilingual and Multi-modal Information Retrieval – Seventh Workshop of the Cross-Language Evaluation Forum, CLEF 2006*, volume 4730 of *LNCS*, pages 595–608, Alicante, Spain, 2007.

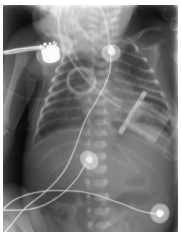
1121-220-454-700
avg. WC: 8.1
train images: 1



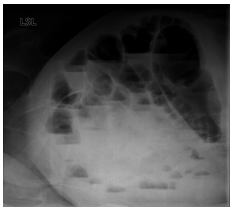
1121-127-800-700
avg. WC: 7.8
train images: 1

1121-220-320-700
avg. WC: 7.7
train images: 1


1121-230-921-700
avg. WC: 7.7
train images: 89



112d-121-500-000
avg. WC: 7.6
train images: 97

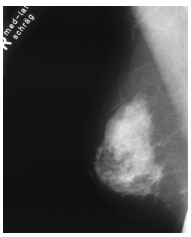


1123-129-700-400
avg. WC: 7.6
train images: 1

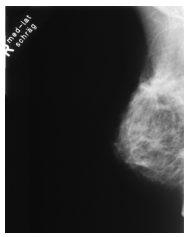
1121-240-442-700
avg. WC: 7.6
train images: 53

1121-210-213-700
avg. WC: 7.5
train images: 15

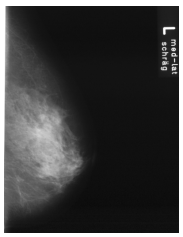
(a) images with most wildcards.



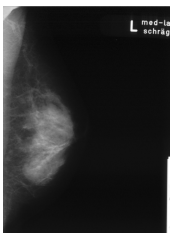
1124-410-610-625
avg. WC: 4.6
train images: 73



1124-410-610-625
avg. WC: 4.5
train images: 73

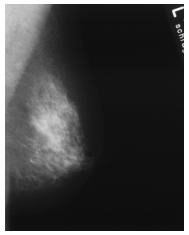


1124-410-620-625
avg. WC: 4.5
train images: 78

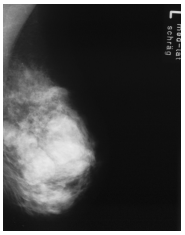


1124-410-620-625
avg. WC: 4.4
train images: 78

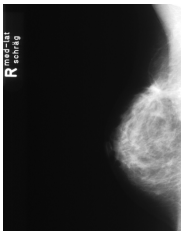
1121-116-917-700
avg. WC: 4.4
train images: 17



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

(b) images with fewest wildcards.

Figure 2: Example images from the IRMA database with their full IRMA code and the average number of wildcards over all runs.