Automatic Medical Image Annotation in ImageCLEF 2007: Overview, Results, and Discussion

Thomas Deselaers [a,*], Thomas M. Deserno [b], and Henning Müller [c,d]

[a] *RWTH Aachen University, Computer Science Department, Aachen, Germany*
[b] *RWTH Aachen University, Department of Medical Informatics, Aachen, Germany*
[c] *University and Hospitals of Geneva, Medical Informatics, Geneva, Switzerland*
[d] *Business Information System, University of Applied Sciences Sierre, Switzerland*

Abstract

In this paper, the automatic medical annotation task of the 2007 CLEF cross-language image retrieval campaign (ImageCLEF) is described. The paper focusses on the images used, the task setup, and the results obtained in the evaluation campaign. Since 2005, the medical automatic image annotation task exists in ImageCLEF with increasing complexity to evaluate the performance of state-of-the-art methods for completely automatic annotation of medical images based on visual properties. The paper also describes the evolution of the task from its origin in 2005 to 2007. The 2007 task, comprising 11,000 fully annotated training images and 1,000 test images to be annotated, is a realistic task with a large number of possible classes at different levels of detail. Detailed analysis of the methods across participating groups is presented with respect to the (i) image representation, (ii) classification method, and (iii) use of the class hierarchy. The results show that methods which build on local image descriptors and discriminative models are able to provide good predictions of the image classes, mostly by using techniques that were originally developed in the machine learning and computer vision domain for object recognition in non-medical images.

Key words: automatic image annotation, medical images, benchmark, evaluation

1. Introduction

Quantitative evaluation of performance is a crucial step in nearly every research and engineering problem. Without quantitative comparison and evaluation of competing approaches it is impossible to determine which directions are promising and which are not. In the past it was shown that evaluation campaigns that independently compare the state-of-the-art systems of different research groups foster improvements (Pallet, 2003) [1].

Centrally organised benchmarks such as the Text REtrieval Conference (TREC) [2] (Voorhees and Harman, 2005) and the NIST Open machine translation evaluation [3] (National Institute of Standards and Technology (NIST), 2001-2008) are well established events. These are organised annually in information retrieval and machine translation, respectively.

The PASCAL Visual Object Classes Challenge (PASCAL VOC) [4], which has been organised annually since 2005, aims at comparing different meth-

* Corresponding author.
Email addresses: deselaers@cs.rwth-aachen.de (Thomas Deselaers), deserno@ieee.org (Thomas M. Deserno), henning.mueller@sim.hcuge.ch (Henning Müller).
URL: www-i6.informatik.rwth-aachen.de/~deselaers (Thomas Deselaers).

[1] http://www.nist.gov/speech/history
[2] http://trec.nist.gov
[3] http://www.nist.gov/speech/tests/mt/index.htm
[4] http://www.pascal-network.org/challenges/VOC

ods for object recognition, detection, and, more recently, segmentation (Everingham, 2006; Everingham et al., 2005). ImagEVAL[5] ran a first evaluation campaign for different aspects of content-based image access in 2006 (Moëllic and Fluhr, 2006). TRECVID[6] is part of TREC and has organised video retrieval evaluations on an annual basis since 2001 with the goal to promote progress in content-based retrieval from digital video. The initiative for the Evaluation of XML Retrieval (INEX)[7] has offered a multimedia track since 2005 with various query and document types.

Furthermore, two technical committees (TCs) of the International Association for Pattern Recognition (IAPR)[8] work on benchmarking and on multimedia systems respectively. The IAPR TC 12[9] actively works on creating the MediaMill challenges (Snoek et al., 2006) and the IAPR TC 5[10] works on benchmarking and software in a more general context in pattern recognition.

ImageCLEF[11] was one of the first campaigns organising evaluation events for image retrieval applications. ImageCLEF is part of the Cross Language Evaluation Forum (CLEF)[12]. CLEF and ImageCLEF are described in Section 2.

The remainder of this paper is structured as follows: Section 2 gives an overview of CLEF with a focus on ImageCLEF and the medical image annotation task. The coding scheme, which is used to represent image annotations, is described in Section 3. The dataset used for the medical image annotation task in ImageCLEF 2007 is described in Section 4. The description of the task is completed with the evaluation scheme that is applied to assess annotation quality in Section 5. In section 6, a short description of the methods that were applied by the individual groups in ImageCLEF 2007 is given and in Section 7, the results of the evaluation are presented. The results are discussed in Section 8, and conclusions are presented in Section 9. In the appendix, we present a table with the results of all runs that were submitted in 2007.

2. CLEF and ImageCLEF

The Cross Language Evaluation Forum[13] (CLEF) originally started as a track for multilingual information access in the Text REtrieval Conference[14] (TREC). It aims at supporting global digital library applications by developing an infrastructure for testing, tuning, and evaluating 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 focuses on multi-lingual and more recently on multi-modal aspects of information retrieval. ImageCLEF 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. They were searched using textual queries written in a different language (Clough and Sanderson, 2004). ImageCLEF 2003 attracted four participants, and the approaches used a range of text-based retrieval and query enhancement techniques such as query expansion. In 2004, a medical and an interactive retrieval task were added to ImageCLEF (Clough et al., 2005). The medical task used a set of images with associated medical case notes and was primarily offered as a query-by-(visual)-example (QBE) retrieval task (Faloutsos et al., 1994) because the search tasks supplied by the organisers contained only images but no text. However, participants could involve text in subsequent retrieval iterations through relevance feedback or query expansion and combine both image processing and text-based retrieval methods. ImageCLEF 2004 attracted participation from 18 research groups across the world, demonstrating the need for such an evaluation campaign. In 2005, a medical image annotation task was added to ImageCLEF and participation increased strongly, in particular for the newly offered image annotation task where 12 groups from 9 countries participated (Clough et al., 2006; Deselaers et al., 2007b). In ImageCLEF 2005 a total of 20 groups participated.

In 2006, the medical annotation task was continued with an enlarged dataset and a higher number of classes, and the database used for medical retrieval grew to approximately 50,000 images (Müller et al., 2007b). The photographic retrieval task used the new IAPR TC 12 database of vacation pho-

[5] http://www.imageval.org
[6] http://www-nlpir.nist.gov/projects/t01v
[7] http://inex.is.informatik.uni-duisburg.de
[8] http://www.iapr.org
[9] http://staff.science.uva.nl/~worring/TC12
[10] http://www.dsic.upv.es/~iaprtc5
[11] http://www.imageclef.org
[12] http://www.clef-campaign.org
[13] http://www.clef-campaign.org/
[14] http://trec.nist.gov/

tographs[15] (Grubinger et al., 2006), and an object detection task was added (Clough et al., 2007). A total of 24 groups participated.

In 2007, 38 groups participated in ImageCLEF. The medical annotation task was extended towards hierarchical classification, the medical retrieval database grew to approximately 70,000 images (Müller et al., 2007a), the photographic retrieval task used sparse textual data (Grubinger et al., 2007), and the object detection task was replaced by an object retrieval task (Deselaers et al., 2007a).

Medical Automatic Image Annotation Tasks 2005 and 2006.

Starting in 2005, automatic medical image annotation has evolved from a simple classification task with about 60 classes to a task with almost 120 classes. From the very start however, it was clear that the number of classes cannot be scaled indefinitely and that the number of classes that are desirable to be recognised in medical applications is far too big to assemble sufficient training data to create suitable classifiers. To address this issue, a hierarchical class structure such as the *Image Retrieval in Medical Applications* (IRMA) code (Lehmann et al., 2003) can be a solution because it supports the creation of a set of classifiers for subproblems.

The classes in the years 2005 and 2006 were based on the IRMA code. They were created by grouping similar codes in a single class. In 2007, the task has changed, and the objective is to predict complete IRMA codes instead of simple classes.

The 2007 medical automatic annotation task builds on top of the task in 2006: 1,000 new images were collected and are used as test data. The training and the test data of 2006 were used as training and development data, respectively.

3. The IRMA Code

Table 1
Example codes for the body region axis.

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 joint
430	upper extremity (arm); forearm
431	upper extremity (arm); forearm; distal forearm
432	upper extremity (arm); forearm; proximal forearm
440	upper extremity (arm); elbow
...	

Existing medical terminologies such as the *Medical Subject Headings* (MeSH) thesaurus are polyhierarchical, 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 of textual classification (Lehmann et al., 2003). In particular, the IRMA code is composed of 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). Some example codes for the body region axis (BBB) are given in Table 1.

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

The potential advantage of using a class hierarchy over using a flat class scheme is that it is in principle possible to create classifiers for large numbers of classes by creating classifiers discriminating between subclasses. Furthermore, a hierarchy-aware classification scheme could potentially be extended when the hierarchy is extended, whereas most flat classification schemes need to be retrained from scratch.

4. Database and Task Description

The complete database consists of 12,000 fully classified medical radiographs taken randomly from clinical routine at the RWTH Aachen University Hospital. 10,000 of these were released along with their classification as training data, another 1,000 were also published with their classification as vali-

[15] http://eureka.vu.edu.au/~grubinger/IAPR/TC12_Benchmark.html

dation data to allow for tuning classifiers in a standardised manner. 1,000 additional images were released at a later date without classification as test data. These 1,000 images had to be classified using the 11,000 images (10,000 training + 1,000 validation) as training data.

Each of the 12,000 images is annotated with its complete IRMA code (see Sec. 3). In total, 116 different IRMA codes occur in the database. The codes are not uniformly distributed, and some codes have a significantly larger share among the data than others (Figure 2). The least frequent codes are represented at least 10 times in the training data to allow for learning suitable models.

Example images from the database together with textual labels and their complete code are given in Figure 1.

5. Hierarchical Classification

To define an evaluation scheme for hierarchical classification, we assume the four axes to be independent and uncorrelated. Hence, we can consider the axes separately and just sum up the errors for each axis individually.

Hierarchical classification is a well-known topic in various fields. The classification of documents is often done using an ontology-based class hierarchy (Sun and Lim, 2001), and in information extraction similar techniques are applied (Maynard et al., 2006). In our case, however, we developed a novel evaluation scheme to account for the particularities of the IRMA code, which considers errors that are made early in a hierarchy to be worse than errors that are made at a fine level, and it is explicitly possible to predict a code partially, i.e. to predict a code up to a certain position and put wild-cards for the remaining positions, which is penalised half as strongly as a misclassification.

Our evaluation scheme is described in the following, where we only consider one axis. The same scheme is applied to each axis individually.

Let $l_1^I = l_1, l_2, \ldots, l_i, \ldots, l_I$ be the *correct* code (for one axis) of an image, i.e. if a classifier predicts this code for an image, the classification is perfect. Further, let $\hat{l}_1^I = \hat{l}_1, \hat{l}_2, \ldots, \hat{l}_i, \ldots, \hat{l}_I$ be the *predicted* code (for one axis) of an image.

The correct code is specified completely: l_i is specified for each position. The classifiers however, are allowed to specify codes only up to a certain level, and predict "*don't know*" (encoded by $*$) for the remaining levels of this axis.

Given an incorrect classification at position $\hat{l}_i$ we consider all succeeding decisions to be wrong and given a non-specified ("don't know") position, we consider all succeeding decisions to be not specified.

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). We can say that a decision at position l_i is correct by chance with a probability of $\frac{1}{b_i}$ if b_i is the number of possible labels (the "branching factor") 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}.

Assembling the ideas from above in a straightforward manner leads to the following equation:

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

with

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

where the parts of the equation account for

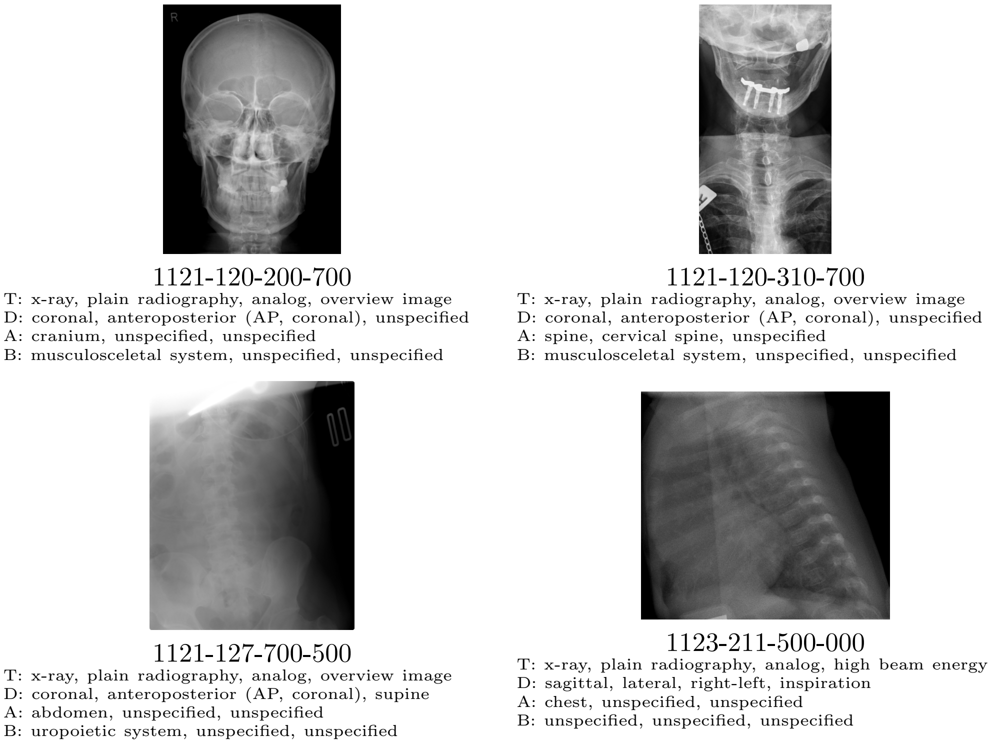
(a) difficulty of the decision at position i (branching factor)

(b) the level in the hierarchy (position in the string)

(c) correct/not specified/wrong, respectively.

In addition, for every code, the maximal possible error is calculated and the errors are normed such that a completely false decision (i.e. all positions false) gets an error count of 1.0 and an in all positions correctly classified image has an error of 0.0.

Table 2 shows examples for a correct code with different predicted codes. Predicting the completely correct code leads to an error measure of 0.0, predicting all positions incorrectly leads to an error measure of 1.0. The examples in Table 2 demonstrate that a classification error in a position to the end of the code results in a lower error measure than a position in one of the first positions. The last column of the table shows the effect of the branching factor b. In this column we assumed $b = 2$ in each node of the hierarchy. It can be observed that the errors



1121-120-200-700
T: x-ray, plain radiography, analog, overview image
D: coronal, anteroposterior (AP, coronal), unspecified
A: cranium, unspecified, unspecified
B: musculosceletal system, unspecified, unspecified

1121-120-310-700
T: x-ray, plain radiography, analog, overview image
D: coronal, anteroposterior (AP, coronal), unspecified
A: spine, cervical spine, unspecified
B: musculosceletal system, unspecified, unspecified

1121-127-700-500
T: x-ray, plain radiography, analog, overview image
D: coronal, anteroposterior (AP, coronal), supine
A: abdomen, unspecified, unspecified
B: uropoietic system, unspecified, unspecified

1123-211-500-000
T: x-ray, plain radiography, analog, high beam energy
D: sagittal, lateral, right-left, inspiration
A: chest, unspecified, unspecified
B: unspecified, unspecified, unspecified

Fig. 1. Example images from the medical annotation task with full IRMA-code and its textual representation.

for the later positions have more weight compared to the real errors in the real hierarchy.

For example, the calculation for the classification 3177 is done as follows:

$$\mathrm{EM}(\text{"}3177\text{"}) = \frac{\frac{1}{10}\cdot\frac{1}{1}\cdot 0+\frac{1}{3}\cdot\frac{1}{2}\cdot 0+\frac{1}{9}\cdot\frac{1}{3}\cdot 1+\frac{1}{16}\cdot\frac{1}{4}\cdot 1}{\frac{1}{10}\cdot\frac{1}{1}\cdot 1+\frac{1}{3}\cdot\frac{1}{2}\cdot 1+\frac{1}{9}\cdot\frac{1}{3}\cdot 1+\frac{1}{16}\cdot\frac{1}{4}\cdot 1}, \qquad (2)$$

where the denominator of the error measure is used to normalise the score according to the maximally possible error. The branching factors for the positions are 10, 3, 9, and 16, respectively, and the individual summands in the nominator and denominator are constructed according to Eq (1).

6. Participating Groups & Methods

In the medical automatic annotation task 2007, 29 groups registered of which 10 groups participated, submitting a total of 68 runs. The group with the highest number of submissions had 30 runs in total.

In the following, groups are listed alphabetically and their methods are described briefly.

Table 2
Example scores for hierarchical classification for one axis. The correct IRMA code is assumed to be TTTT = 318a. The columns denote (from left to right) hypothesised codes, the error measure as described above, and the error measure where a branching factor $b = 2$ is assumed in each node in the hierarchy.

classified	error measure	error measure (b=2)
318a	0.000	0.000
318*	0.024	0.060
3187	0.049	0.120
31*a	0.082	0.140
31**	0.082	0.140
3177	0.165	0.280
3***	0.343	0.260
32**	0.687	0.520
1000	1.000	1.000

6.1. *BIOMOD: University of Liege, Belgium*

The Bioinformatics and Modelling group from the University of Liege [16] in Belgium submitted four

[16] http://www.montefiore.ulg.ac.be/services/stochastic/biomod

runs. The approach is based on an object recognition framework using extremely randomised trees and randomly extracted sub-windows (Marée et al., 2005). All runs use the same technique but differ in the way the code is assembled. One run predicts the full code, one run predicts each axis independently and the other two runs are combinations of these.

6.2. *BLOOM: IDIAP, Switzerland*

The Blanceflor-om2-toMed group from IDIAP in Martigny, Switzerland submitted 7 runs. All runs use support vector machines (either in one-against-one or one-against-the-rest manner). Features used are downscaled versions of the images, SIFT (Scale-Invariant Feature Transform) features extracted from sub-images, and combinations of these (Tommasi et al., 2007).

6.3. *GENEVA: medGIFT Group, Switzerland*

The medGIFT group[17] from Geneva, Switzerland submitted 3 runs, each of the runs uses the GIFT (GNU Image Finding Tool) image retrieval system. Different voting strategies were used to obtain classifications at different depths of the code hierarchy (Zhou et al., 2007).

6.4. *CYU: Information Management AI lab, Taiwan*

The Information Management AI lab from the Ching Yun University of Jung-Li, Taiwan submitted one run using a nearest neighbour classifier using different global and local image features which are particularly robust with respect to lighting changes.

6.5. *MIRACLE: Madrid, Spain*

The Miracle group from Madrid, Spain[18] submitted 30 runs. The classification was done using a 10-nearest neighbour classifier and the features used are gray-value histograms, Tamura texture features, global texture features, and Gabor features, which were extracted using FIRE. The runs differ in the features used, how the prediction was done (predicting the full code, axis-wise prediction, different subsets of axes jointly), and whether the features were normalised or not.

6.6. *OHSU: Oregon Health and Science University, Portland, OR, USA*

The Department of Medical Informatics and Clinical Epidemiology[19] of the Oregon Health and Science University in Portland, Oregon submitted two runs using neural networks and GIST descriptors. One of the runs uses a support vector machine as a second level classifier to help in discriminating the two most difficult classes.

6.7. *RWTHi6: RWTH Aachen University, Aachen, Germany*

The Human Language Technology and Pattern Recognition group[20] of the RWTH Aachen University in Aachen, Germany submitted 6 runs; all are based on sparse histograms of image patches, which were obtained by extracting patches at each position in the image. The histograms have 65536 or 4096 bins (Deselaers et al., 2006). The runs differ in the resolution of the images. One run is a combination of 4 normal runs, and one run does the classification axis-wise. The other runs directly predict the full code.

6.8. *IRMA: RWTH Aachen University, Medical Informatics, Aachen, Germany*

The IRMA (Image Retrieval for Medical Applications) group from the RWTH Aachen University Hospital[21], in Aachen, Germany submitted three baseline runs using weighted combinations of nearest neighbour classifiers using texture histograms, image cross correlations, and the image deformation model. The parameters used are exactly the same as used in previous years. The runs differ in the way in which the codes of the five nearest neighbours are used to assemble the final predicted code.

[17] http://www.sim.hcuge.ch/medgift

[18] http://www.mat.upm.es/miracle/introduction.html

[19] http://www.ohsu.edu/dmice

[20] http://www-i6.informatik.rwth-aachen.de

[21] http://www.irma-project.org

6.9. *UFR: University of Freiburg, Computer Science Dep., Freiburg, Germany*

The Pattern Recognition and Image Processing group from the University Freiburg[22], Germany, submitted four runs using relational features calculated around interest points which are later combined to form cluster cooccurrence matrices (Setia et al., 2006). Three different classification methods were used: a flat classification scheme using all of the 116 classes, an axiswise-flat classification scheme (i.e. 4 multi-class classifiers), and a binary classification tree (BCT) based scheme. The BCT based approach is much faster to train and classify, but comes at a slight performance penalty. The tree was generated as described in (Setia and Burkhardt, 2007).

6.10. *UNIBAS: University of Basel, Switzerland*

The Databases and Information Systems group from the University of Basel[23], Switzerland submitted 14 runs using a pseudo two-dimensional hidden Markov model to model image deformation in the images that were scaled down, keeping the aspect ratio such that the longer side has a length of 32 pixels (Springmann and Schuldt, 2007). The runs differ in the features (pixels, Sobel features) that were used to determine the deformation and in the k-parameter for the k-nearest neighbour classifier.

7. Results

The results of the evaluation are given in Table 3 ordered by group. A full list of all submitted runs is also given in Appendix A. Table 3 gives for each group the number of submitted runs, the best and the worst rank, as well as the minimum, maximum, mean, and median error count and classification error rate. The groups are ordered by the error score of their best submission and it can be seen that there are three groups of submissions: groups with a best error count of approximately 30, groups with an error score between 30 and 80, and groups with worse results.

The method that had the best result in 2006 is at rank 8 in 2007. The method with the best result in 2005 is the main component of the runs on ranks 17 to 25 in 2007. This gives a sense of how much improvement in this field has been achieved since 2005.

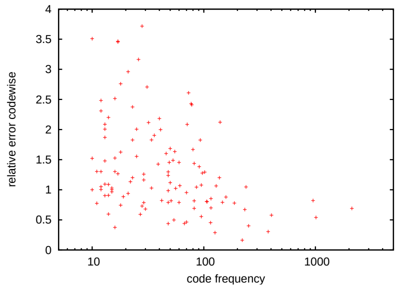


Fig. 3. Code-wise relative error as a function of the frequency of this code in the training data.

8. Discussion

Figure 2 (bottom) is the average confusion matrix over all submitted runs, with the correct class on the y-axis and the predicted class on the x-axis. The 13 columns at the right border of the confusion matrix denote classifications, with 1 to 13 (from left to right) wildcards. That is, the right-most column denotes classifications where no single code position was predicted but each position was unspecified. The classes in the confusion matrix are sorted by frequency of the class in the training data. The frequency of the classes in the training data is plotted in the upper part of Figure 2. The most outstanding feature of the confusion matrix is that a large portion of the images are classified correctly on the average. Furthermore, it can be observed that due to the skewed class distribution to the low class numbers, there are hardly any misclassifications from frequent classes to more rare classes but only from rare classes (high class number) to frequent classes (low class number). This effect can be explained by the higher prior probabilities for the more frequent classes.

The matrix also shows that the classes which are well represented in the training data are more likely to be classified correctly. Figure 3 directly shows the connection between classification error and amount of training data. The x-axis of Figure 3 gives the frequency of the classes/codes in the training data and the y-axis gives the relative error for the codes averaged over all submitted runs. It can be observed that classes that occur rarely in the training data are more likely to have high errors (top left region), whereas frequent classes are seldom misclassified.

Analysing the results for individual images, we

[22] http://lmb.informatik.uni-freiburg.de
[23] http://dbis.cs.unibas.ch

Table 3
Results of the evaluation by participating group. For each group, the number of submitted runs, the rank of the best and worst run, and the minimum, maximum, mean, and medium error count and error rate are given.

		rank		score				ER			
group	submissions	min	max	min	max	mean	median	min	max	mean	median
BIOMOD	4	30	35	73.82	95.25	80.90	77.26	22.90	36.00	29.28	29.10
BLOOM	7	1	29	26.85	72.41	40.44	29.46	10.30	20.80	13.77	11.50
GENEVA	3	63	65	375.72	391.02	385.68	390.29	99.00	99.70	99.33	99.30
CYU	1	33	33	79.30	79.30	79.30	79.30	25.30	25.30	25.30	25.30
MIRACLY	30	36	68	158.82	505.62	237.42	196.18	49.30	89.00	62.09	55.50
OHSU	2	26	27	67.81	67.98	67.89	67.89	22.70	22.70	22.70	22.70
RWTHi6	6	6	13	30.93	44.56	35.16	33.88	11.90	17.80	13.38	12.55
IRMA	3	17	34	51.34	80.47	61.45	52.54	18.00	45.90	27.97	20.00
UFR	5	7	16	31.44	48.41	41.29	45.48	12.10	17.90	15.36	16.80
UNIBAS	7	19	25	58.15	65.09	61.64	61.41	20.20	23.20	22.26	22.50

noted that only one image was classified correctly by all submitted runs (top left image in Fig. 1). No image was misclassified by all runs. The image which was misclassified most frequently has an average error score of 0.6 over all runs.

Analysing the results, it can be observed that the top-performing runs do not consider the hierarchical structure of the given task, but rather use each individual code as one class and train a 116-class classifier. This approach seems to work best given the currently limited amount of codes, but obviously would not scale up indefinitely and would probably lead to a very high demand for appropriate training data if a much larger amount of classes is to be distinguished. The best run using the hierarchy is on rank 6. It builds on top of the other runs from the same group and uses the hierarchy only in a second stage to combine the four runs.

One common way to achieve improvements is to combine several runs. After the evaluation was over, we combined the best runs of the top 3 groups (BLOOM/IDIAP, RWTH Aachen University, and UFR) using a voting scheme, where a wildcard is set whenever the runs disagree about a particular position. This results in an error score of 24 (error rate of 10.3), which shows that using the code to combine runs can lead to an improvement of the score, but not of the error rate as every code which includes a wildcard is misclassified. This resulting run uses a total of 52 wildcards on 31 images.

Furthermore, it can be seen that if a method is applied that accounts for the hierarchy/axis structure of the code and if a second method is applied that uses the straightforward classification, the latter one outperforms the first (see the runs on ranks 11 and 13 as well as the runs on ranks 7 and 14, 16).

Another clear observation is that methods using local image descriptors outperform methods using global image descriptors. In particular, the top 16 runs all use either local image features alone or local image features in combination with a global descriptor. The runs on the ranks 17-25 use local features to obtain deformation fields to compare the images globally, and the runs on rank 26 and 27 are the best runs using pure global image descriptors.

Considering the ranking with respect to the applied hierarchical measure and the ranking with according to the error rate it becomes obvious that there are hardly any differences. Most of the differences are clearly due to use of the code (mostly inserting of wildcard characters) which can lead to an improvement for the hierarchical evaluation scheme, but will always lead to a deterioration of the error rate.

9. Conclusion

The progression of the ImageCLEF medical automatic annotation tasks from 2005 to 2007 clearly shows that the image recognition community needs evaluation campaigns like ImageCLEF where specialised methods as well as general purpose image recognition and machine learning techniques can be applied and compared based on the same grounds. In 2005, the rather simple task drew a lot of interest and some groups participated in each year. The task

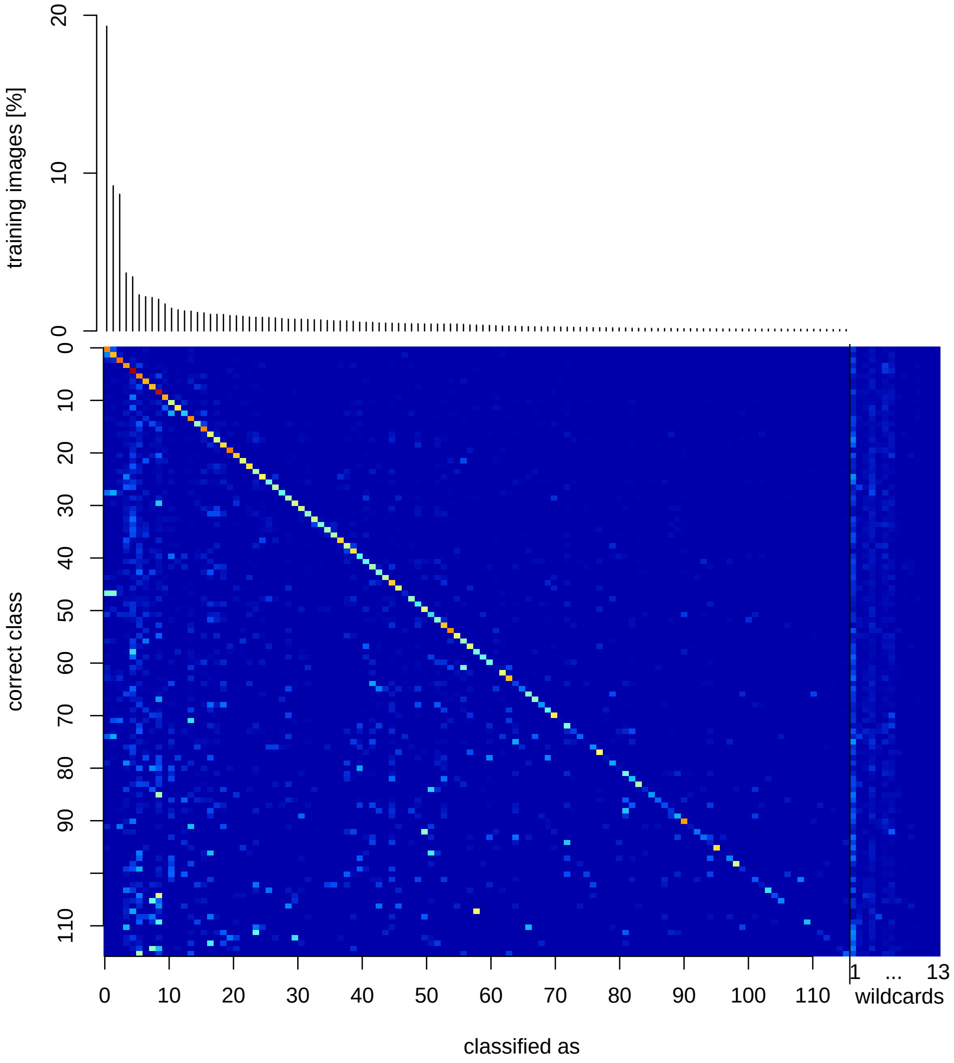


Fig. 2. Confusion matrix and relative frequency of classes in training data.

was continued with increasing complexity in 2006 and 2007.

The task is now at the point where it can be applied directly to images being inserted into a medical picture archiving system. Now, the question arises whether further evaluations for this type of task are required in the future. The main problem in the 2007 task is that it did not force participants to use the hierarchical class structure, which would be a requirement if the classes spanned the whole hierarchy, since it is not feasible to produce sufficient training data to create flat classifiers for such a high number of classes.

For the ImageCLEF 2008 evaluation we plan to extend the task toward using more classes with only little support in the training data, to force participants to use wildcards in their classifications.

References

Clough, P., Grubinger, M., Deselaers, T., Hanbury, A., Müller, H., 2007. Overview of the ImageCLEF 2006 photographic retrieval and object annotation tasks. In: Evaluation of Multilingual and Multi-modal Information Retrieval – Seventh Workshop of the Cross-Language Evaluation Forum, CLEF 2006. Vol. 4730 of Lecture Notes in Computer Series. Alicante, Spain, pp. 579–594.

Clough, P., Müller, H., Deselaers, T., Grubinger, M., Lehmann, T., Jensen, J., Hersh, W., 2006. The CLEF 2005 cross-language image retrieval track. In: Accessing Multilingual Information Repositories, 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005. Vol. 4022 of Lecture Notes in Computer Science. Vienna, Austria, pp. 535–557.

Clough, P., Müller, H., Sanderson, M., 2005. The CLEF Cross Language Image Retrieval Track (ImageCLEF) 2004. In: Fifth Workshop of the Cross–Language Evaluation Forum (CLEF 2004). Vol. 3491 of LNCS. pp. 597–613.

Clough, P. D., Sanderson, M., Nov. 2004. The clef 2003 cross language image retrieval track. In: Comparative Evaluation of Multilingual Information Access Systems. Vol. 3237 of LNCS. Trondheim, Norway, pp. 581–593.

Deselaers, T., Hanbury, A., Viitaniemi, V., Benczúr, A., Brendel, M., Daróczy, B., Escalante Balderas, H. J., Gevers, T., Hern'andez Gracidas, C. A., Hoi, S. C. H., Laaksonen, J., Li, M., Marin Castro, H. M., Ney, H., Rui, X., Sebe, N., Stöttinger, J., Wu, L., Sep. 2007a. Overview of the ImageCLEF 2007 object retrieval task. In: Working notes of the CLEF 2007 Workshop. Budapest, Hungary.

Deselaers, T., Hegerath, A., Keysers, D., Ney, H., Sep. 2006. Sparse patch-histograms for object classification in cluttered images. In: DAGM 2006, Pattern Recognition, 27th DAGM Symposium. Vol. 4174 of Lecture Notes in Computer Science. Berlin, Germany, pp. 202–211.

Deselaers, T., Müller, H., Clough, P., Ney, H., Lehmann, T. M., Aug. 2007b. The CLEF 2005 automatic medical image annotation task. International Journal of Computer Vision 74 (1), 51–58.

Everingham, M., Gool, L. V., Williams, C., Zisserman, A., Apr. 2005. Pascal visual ob ject classes challenge results. Tech. rep., University of Oxford, Oxford, UK.

Everingham, M. e. a., 2006. The 2005 pascal visual object classes challenge. In: Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment (PASCAL Workshop 05). No. 3944 in Lecture Notes in Artificial Intelligence. Southampton, UK, pp. 117–176.

Faloutsos, C., Barber, R., Flickner, M., Hafner, J., Niblack, W., Petkovic, D., Equitz, W., Jul. 1994. Efficient and effective querying by image content. Journal of Intelligent Information Systems 3 (3/4), 231–262.

Grubinger, M., Clough, P., Hanbury, A., Müller, H., Sep. 2007. Overview of the imageclefphoto 2007 photographic retrieval task. In: Working notes of the CLEF 2007 Workshop. Budapest, Hungary.

Grubinger, M., Clough, P., Müller, H., Deselaers, T., May 2006. The IAPR benchmark: A new evaluation resource for visual information systems. In: LREC 06 OntoImage 2006: Language Resources for Content-Based Image Retrieval. Genoa, Italy.

Lehmann, T. M., Schubert, H., Keysers, D., Kohnen, M., Wein, B. B., 2003. The irma code for unique classification of medical images. In: Proceedings SPIE. No. 5033. pp. 440–451.

Marée, R., Geurts, P., Piater, J., Wehenkel, L., June 2005. Random subwindows for robust image classification. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2005). Vol. 1. IEEE, San Diego, CA, USA, pp. 34–40.

Maynard, D., Peters, W., Li, Y., 2006. Metrics for evaluation of ontology-based information extraction. In: Evaluation of Ontologies for the Web. Edinburgh, UK.

Moëllic, P.-A., Fluhr, C., 2006. ImageEVAL 2006 official campaign. Tech. rep., ImagEVAL.

Müller, H., Deselaers, T., Kim, E., Kalpathy-Kramer, J., Deserno, T. M., Hersh, W., Sep. 2007a. Overview of the ImageCLEFmed 2007 medical retrieval and annotation tasks. In: Working notes of the CLEF 2007 Workshop. Budapest, Hungary.

Müller, H., Deselaers, T., Lehmann, T., Clough, P., Hersh, W., 2007b. Overview of the ImageCLEFmed 2006 medical retrieval and annotation tasks. In: Evaluation of Multilingual and Multimodal Information Retrieval – Seventh Workshop of the Cross-Language Evaluation Forum, CLEF 2006. Vol. 4730 of LNCS. Alicante, Spain, pp. 595–608.

National Institute of Standards and Technology (NIST), 2001-2008. NIST open MT machine translation evaluation, http://www.nist.gov/speech/tests/mt/index.htm.

Pallet, D. S., 2003. A look at NIST's benchmark asr tests: Past, present, and future. Tech. rep., National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA.
URL http://www.nist.gov/speech/history/

Setia, L., Burkhardt, H., 2007. Learning taxonomies in large image databases. In: ACM SIGIR Workshop on Multimedia Information Retrieval. Amsterdam, Holland.

Setia, L., Teynor, A., Halawani, A., Burkhardt, H., 2006. Image classification using cluster-cooccurrence matrices of local relational features. In: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval. Santa Barbara, CA, USA.

Snoek, C. G., Worring, M., van Gemert, J. C., Geusebroek, J.-M., Smeulders, A. W., Oct. 2006. The challenge problem for automated detection of 101 semantic concepts in multimedia. In: ACM Multimedia. Santa Barbara, CA, USA, pp. 421–430.

Springmann, M., Schuldt, H., Sep. 2007. Speeding up idm without degradation of retrieval quality. In: Working Notes of the CLEF Workshop 2007.

Sun, A., Lim, E.-P., Nov. 2001. Hierarchical text classification and evaluation. In: IEEE International Conference on Data Mining (ICDM 2001). San Jose, CA, USA, pp. 521–528.

Tommasi, T., Orabona, F., Caputo, B., Sep. 2007. CLEF2007 Image Annotation Task: an SVM–based Cue Integration Approach. In: Working Notes of the 2007 CLEF Workshop. Budapest, Hungary.

Voorhees, E. M., Harman, D. K., 2005. TREC: Experiment and Evaluation in Information Retrieval (Digital Libraries and Electronic Publishing). The MIT Press.

Zhou, X., Gobeill, J., Ruch, P., Müller, H., Sep. 2007. University and Hospitals of Geneva at ImageCLEF 2007. In: Working Notes of the 2007 CLEF Workshop. Budapest, Hungary.

Appendix A. Results of all runs

Table A.1
Results of the medical image annotation task. Score is the hierarchical evaluation score, and ER is the error rate in % that was used in 2005 and 2006 to evaluate the annotation results.

rank	run id	score	ER
1	BLOOM-BLOOM_MCK_oa	26.8	10.3
2	BLOOM-BLOOM_MCK_oo	27.5	11.0
3	BLOOM-BLOOM_SIFT_oo	28.7	11.6
4	BLOOM-BLOOM_SIFT_oa	29.5	11.5
5	BLOOM-BLOOM_DAS	29.9	11.1
6	RWTHi6-4RUN-MV3	30.9	13.2
7	UFR-UFR_cooc_flat	31.4	12.1
8	RWTHi6-SH65536-SC025-ME	33.0	11.9
9	UFR-UFR_cooc_flat2	33.2	13.1
10	RWTHi6-SH65536-SC05-ME	33.2	12.3
11	RWTHi6-SH4096-SC025-ME	34.6	12.7
12	RWTHi6-SH4096-SC05-ME	34.7	12.4
13	RWTHi6-SH4096-SC025-AXISWISE	44.6	17.8
14	UFR-UFR_cooc_codewise	45.5	17.9
15	UFR-UFR_cooc_tree2	47.9	16.9
16	UFR-UFR_cooc_tree	48.4	16.8
17	rwth_mi_k1_tn9.187879e-05_common.run	51.3	20.0
18	rwth_mi_k5_majority.run	52.5	18.0
19	UNIBAS-DBIS-IDM_HMM_W3_H3_C	58.1	22.4
20	UNIBAS-DBIS-IDM_HMM2_4812_K3	59.8	20.2
21	UNIBAS-DBIS-IDM_HMM2_4812_K3_C	60.7	23.2
22	UNIBAS-DBIS-IDM_HMM2_4812_K5_C	61.4	23.1
23	UNIBAS-DBIS-IDM_HMM2_369_K3_C	62.8	22.5
24	UNIBAS-DBIS-IDM_HMM2_369_K3	63.4	21.5
25	UNIBAS-DBIS-IDM_HMM2_369_K5_C	65.1	22.9
26	OHSU-OHSU_2	67.8	22.7
27	OHSU-gist_pca	68.0	22.7
28	BLOOM-BLOOM_PIXEL_oa	68.2	20.1
29	BLOOM-BLOOM_PIXEL_oo	72.4	20.8
30	BIOMOD-full	73.8	22.9
31	BIOMOD-correction	75.8	25.3
32	BIOMOD-safe	78.7	36.0
33	im.cyu.tw-cyu_w1i6t8	79.3	25.3
34	rwth_mi_k5_common.run	80.5	45.9
35	BIOMOD-independant	95.3	32.9
36	miracle-miracleAAn	158.8	50.3
37	miracle-miracleVAn	159.5	49.6
38-60	runs from miracle group	–	
61	miracle-miracleVA	325.9	85.2
62	miracle-miracleVATABD	350.2	89.0
63	GE-GE_GIFT10_0.5ve	375.7	99.7
64	GE-GE_GIFT10_0.15vs	390.3	99.3
65	GE-GE_GIFT10_0.66vd	391.0	99.0
66	miracle-miracleVATDAB	419.7	84.4
67	miracle-miracleVn	490.7	82.6
68	miracle-miracleV	505.6	86.8