Overview of the CLEF 2009 Medical Image Annotation Track

Tatiana Tommasi[1], Barbara Caputo[1], Petra Welter[2], Mark Oliver Güld[2], and Thomas M. Deserno[2]

[1] Idiap Research Institute, Martigny, Switzerland,
{ttommasi, bcaputo}@idiap.ch
[2] RWTH, Aachen University,
Dept. of Medical Informatics, Aachen, Germany
{pwelter, güld, tdeserno}@mi.rwth-aachen.de

Abstract. This paper describes the last round of the medical image annotation task in ImageCLEF 2009. After four years, we defined the task as a survey of all the past experience. Seven groups participated to the challenge submitting nineteen runs. They were asked to train their algorithms on 12677 images, labelled according to four different settings, and to classify 1733 images in the four annotation frameworks. The aim is to understand how each strategy answers to the increasing number of classes and to the unbalancing. A plain classification scheme using support vector machines and local descriptors outperformed the other methods.

1 Introduction

The medical image annotation task was introduced in the ImageCLEF[3] challenge in 2005. Its main contribution was to provide a resource for benchmarking content-based image classification systems focusing on medical images. Hospitals collect hundreds of imaging data everyday and automatic image annotation can be an important step when searching for images in huge databases. Automatic techniques able to identify acquisition modality, body orientation, body region, and biological system examined could be used for multilingual image annotations as well as for DICOM header corrections in medical image acquisition routine.

Over the last four years the medical annotation task evolved both in terms of number of images, classes, and classes' framework provided. It was born as a 60 plain class problem [3], grew up to a 120 class problem [6], and became a complex hierarchical class task in 2007 [5, 2]. In 2008, class imbalance was added to foster the use of prior knowledge encoded into the hierarchy of classes [1].

This year we celebrate the 5th medical image annotation task anniversary and we decided to organize its conclusive round as a survey on the last years experience. The idea is to compare the scalability of different image classification techniques as the number of classes grows, their hierarchical structure increase, and badly populated classes appear.

[3] http://www.imageclef.org/

2 Database and Task Description

As in the past challenge editions, the annotation task was defined on the basis of the IRMA project[4]. This year a database of 12677 fully classified radiographs, taken randomly from medical routine, was made available as training set. Images are labelled according to four classification label sets considering:

- 57 classes as in 2005 (12631 images) + a "clutter" class C (46 images);
- 116 classes as in 2006 (12334 images) + a "clutter" class C (343 images);
- 116 IRMA codes as in 2007 (12334 images) + a "clutter" class C (343 images);
- 193 IRMA codes as in 2008 (12677 images).

For the first two label settings, images are associated to simple raw numbers while in the last two label settings images are identified by their complete IRMA code (see Section 3). The 1-57 labels used for the first group definition are derived through a high level identification of images in IRMA code terms. Considering a more detailed image annotation and the introduction of some new classes we pass to 116 and then to 193 classes. The "clutter" class for a specific setting contains all the images belonging to new classes, or images described with a higher level of detail in the final 2008 setting.

The test data consisted of 1733 images. Not all the training classes have examples in this set:

- 2005 labelling - 55 classes (out of 57) with 1639 images + class C with 94 images;
- 2006 labelling - 109 classes (out of 116) with 1353 images + class C with 380 images;
- 2007 labelling - 109 IRMA codes (out of 116) with 1353 images + class C with 380 images;
- 2008 labelling - 169 IRMA codes (out of 193) with 1733 images.

Note the distribution of the images in the classes of the training set: for 2005, 2006 and 2007 classes have more than 6 images while in 2008 there are classes with 1 to 5 images. Concerning the 2008 labels, the test data have a 20% of images which are badly (classes with less than 10 images) represented in the training data.

Participants to the medical annotation task were asked to classify the test images according to all the four label settings. Each group is allowed to submit different runs, but each of them should be based only on one algorithm which should be optimized to face the four different classification problems. The aim is to understand how each algorithm answers to the increasing number of classes and to the unbalancing. The classification results are considered per year and the error scores are summed to have a final unique way to rank the performance of the submitted runs.

[4] http://irma-project.org/index_en.php

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); elbow
...	

Table 1. Examples from the IRMA code

3 IRMA Code

Standardized nomenclature for medical imaging are generally roughly structured, ambiguous, and often use optional tags. Concerning the needs for content-based image retrieval and annotation in the medical field, a detailed unambiguous coding scheme is required. Valid relations between code and sub-code elements could be "is-a" and "part-of", defining a strict hierarchical order. Causality is also important for grouping of processing strategies. Therefore, a mono-hierarchical scheme is required, where each sub-code element is connected to only one code element. Since categorization of medical images must cover all aspects influencing the image content and structure, a multi-axial scheme is needed [4].

The IRMA code strictly rely on these rules. It is composed from four axes having three to four positions each in {0, ..., 9, a, ..., z}, where "0" denotes "unspecified " to determine the end of a path along an axis:

- the technical code (T) describes the image modality;
- the directional code (D) models body orientations;
- the anatomical code (A) refers to the body region examined;
- 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 image modalities are introduced. Based on the hierarchy, the more code position differ from "0", the more detailed is the description.

classified	error score
18	0.0
21	1.0
*	0.5

Table 2. Error score evaluation for 2005 and 2006 settings. The correct label is 18.

4 Error Evaluation

We describe here how the error score for the medical image annotation challenge was evaluated. On the basis of the image labelling, we defined two different evaluation strategies.

2005 and 2006. For these two years the error is evaluated just on the capability of the algorithm to make the correct decision. There is also the possibility to say *"don't know"*, which is encoded by "*". An example is given in Table 2.

2007 and 2008. For these two years, the error is evaluated on the basis of the hierarchical IRMA code.

Let an image be coded by the technical, directional, anatomical and biological independent axes. They can be considered separately and we can 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 every position, and in $\hat{l}_i$ 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.

We want to penalize 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, 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 for position i. This assumes equal priors for each class at each position.

Furthermore, we want to penalize 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

classified	error count
463	0.000000
46*	0.025531
461	0.051061
4*1	0.069297
4**	0.069297
47*	0.138594
473	0.138594
477	0.138594
**	0.125000
731	0.250000

Table 3. Error score evaluation for 2007 and 2008 settings. We are considering just one axis, the correct label is 463.

together:

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

with

$$\delta(l_i, \hat{l_i}) = \begin{cases} 0 \text{ if } l_j = \hat{l_j} \ \forall j \leq i \\ 0.5 \text{ if } l_j = * \ \exists j \leq i \\ 1 \text{ if } l_j \neq \hat{l_j} \ \exists j \leq i \end{cases} \tag{2}$$

where the parts of the equation:

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

In addition, for every axis, the maximal possible error is calculated and the errors are normalized 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. Finally setting a wildcard "*" instead of a "0" is not considered a mistake (see Table 3).

Clutter in 2005, 2006 and 2007. For these three years we introduced a class called "clutter" C. Even if in the test set there are images belonging to this class, their classification do NOT influence the error score for the challenge (see Table 4). An example of the released database complete labelling is in Figure 1.

5 Participation

In 2009, seven groups participated in the medical annotation task submitting nineteen runs in total. In the following we describe the methods applied by the participating groups.

classified 2005-06	error count
18	0.0
21	0.0
*	0.0
C	0.0
classified 2007	error count
111	0.000000
11*	0.000000
1**	0.000000
***	0.000000
C	0.000000

Table 4. Error score evaluation for the clutter class.The correct label is C or CCC.

TAUbiomed. The Medical Image Processing Lab from Tel Aviv University in Israel submitted one run using a multiple-resolution patch-based bag-of-visual words approach. Classification is performed through support vector machines. The code hierarchy is completely neglected and no wildcards "*" were used.

Idiap. The Idiap Research Institute from Switzerland submitted four runs reproposing the same strategies used in 2008. They consisted in different classification schemes for support vector machines coupling two different image descriptors.

FEITIJS. The Faculty of Electrical Engineering and Information Technologies from the University of Skopje in Macedonia submitted one run. It is based on global and local image descriptors which are classified using bagging and random forest.

VPA. The Computer Vision and Pattern Analysis Laboratory from Sabanci University in Turkey submitted five runs. They used local binary patterns as features and support vector machine as classifier. They adopted a hierarchical approach considering, when applicable, the four IRMA code axes separately.

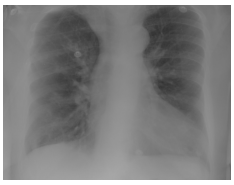
medGIFT. The medGIFT group from University Hospitals of Geneva in Switzerland submitted three runs using different descriptors and voting schemes in the medGIFT image retrieval system.

DEU. The Dokuz Eylul University in Turkey participated submitting four runs. Different global and local features are extracted from images and classification is performed with a k-Nearest Neighbour algorithm.

IRMA. As a general reference the Image Retrieval in Medical Application group at RWTH Aachen University, Germany, provided a baseline run. It was defined using Tamura Texture Measures and the Image Distortion Model. Since 2004, the parametrization is unchanged, so the IRMA code hierarchy is disregarded.

The results of the challenge evaluation are given in Table 5, sorted by error score sum over the four year label settings. Considering the error score per-year,

2005: 22 (11-4-91-7)
2006: 54
2007: 1121-4a0-914-700
2008: 1121-4a0-914-700



2005: 1 (11-1-50-0)
2006: 1
2007: 1123-127-500-000
2008: 1123-127-500-000

2005: 50 (11-2-45-7)
2006: C
2007: CCCC-CCC-CCC-CCC
2008: 1121-230-451-700

2005: C
2006: C
2007: CCCC-CCC-CCC-CCC
2008: 1127-310-600-625

Fig. 1. Examples of the four years labels settings.

the group ranking does not change except for an exchange of the first and second rank positions between the Idiap and TAU group in 2006.

In general, analyzing the results it can be seen that the top-performing runs do not consider the hierarchical structure of the given task (2007 and 2008 labels), but rather use each individual code as one class and train a plain classifier.

Comparing the 2005 and 2006 results, we see that there is a general decrease in the error score. A possible explanation is that in 2005 the 57 classes are wide, each one containing different sublevels in terms of IRMA codes. This make them difficult to be modelled by a classifier in the training phase. On the other hand, comparing the 2007 and 2008 results there is a general increase in the error score. This effect was expected: here new classes with the same level of detail respect to the IRMA code are added passing from 2007 to 2008. Moreover some of the new classes are poorly populated in the training set.

As final remark, we notice that methods using patch-based local image descriptors and discriminative SVM classification methods outperform the other approaches.

Run & error score	2005	2006	2007	2008	SUM
TAUbiomed_95_9_1246120389711	356	263	64.3	169.5	852.8
Idiap_3_9_1245417716666	393	260	67.23	178.93	899.16
Idiap_3_9_1245417533955	393	260	67.23	179.17	899.4
Idiap_3_9_1245417469975	447	292	75.81	224.82	1039.63
Idiap_3_9_1245417671272	447	292	75.81	227.19	1042
FEITIJS_96_9_1245937057229	549	433	128.10	242.46	1352.56
VPA SabanciUniv_63_9_1245419336923	578	462	155.05	261.16	1456.21
VPA SabanciUniv_63_9_1245418900571	578	462	201.31	272.61	1513.92
VPA SabanciUniv_63_9_1245944101876	587	498	169.33	300.44	1554.77
VPA SabanciUniv_63_9_1246033855761	587	502	172.08	320.61	1581.69
MedGIFT_77_9_1245961041705	618	507	190.73	317.53	1633.26
MedGIFT_77_9_1245971471117	618	507	190.73	317.53	1633.26
IRMA	790	638	207.55	359.29	1994.84
MedGIFT_77_9_1246044416990	791.5	612.5	272.69	420.91	2097.6
VPA SabanciUniv_63_9_1245936277557	587	1170	413.1	574	2744.1
DEU_97_9_1246226037987	1368	1183	487.5	642.5	3681
DEU_97_9_1246225040330	1370	1189	488.5	639	3686.5
DEU_97_9_1245952497879	1471	1243	541.8	713.3	3969.1
DEU_97_9_1245952673253	1484	1246	539.7	710.1	3979.8

Table 5. Results from the medical image annotation task.

5.1 Discussion and Conclusion

We have presented the ImageCLEF 2009 medical image annotation task. This is its conclusive round and we organized it as a survey on the last four years experience. We want to compare the scalability of different image classification techniques as the number of classes grows, their hierarchical structure increase, and badly populated classes appear. A plain classification scheme using support vector machine and local descriptors outperformed the other methods. The obtained scores range from 852.8, over 1994.84, to 3979.8 for best, baseline and worst respectively.

6 Acknowledgements

We would like to thank the CLEF campaign for supporting the ImageCLEF initiative. The authors T. Tommasi and B. Caputo are supported by the EMMA project thanks to the Hasler foundation (www.haslerstiftung.ch). P. Welter is supported by the German Research Foundation (DFG, Le 1108/9).

References

1. Thomas Deselaers and Thomas M. Deserno. Medical image annotation in ImageCLEF 2008. In Carol Peters, Danilo Giampiccolo, Nicola Ferro, Vivien Petras, Julio

Gonzalo, Anselmo Peñas, Thomas Deselaers, Thomas Mandl, Gareth Jones, and Mikko Kurimo, editors, *Evaluating Systems for Multilingual and Multimodal Information Access — 9th Workshop of the Cross–Language Evaluation Forum*, Lecture Notes in Computer Science, Aarhus, Denmark, September 2009 – to appear.

2. Thomas Deselaers, Thomas M. Deserno, and Henning Müller. Automatic medical image annotation in ImageCLEF 2007: Overview, results, and discussion. *Pattern Recognition Letters*, 29(15):1988–1995, 2008.
3. Thomas Deselaers, Henning Müller, Paul Clough, Hermann Ney, and Thomas M. Lehmann. The CLEF 2005 automatic medical image annotation task. *International Journal in Computer Vision*, 74(1):51–58, 2007.
4. Thomas M. Lehmann, Henning Schubert, Daniel Keysers, Michael Kohnen, and Berthold B. Wein. The IRMA code for unique classification of medical images. In H. K. Huang and O. M. Ratib, editors, *Medical Imaging 2003: PACS and Integrated Medical Information Systems: Design and Evaluation.*, volume 5033 of *SPIE Proceedings*, pages 440–451, San Diego, California, USA, May 2003.
5. Henning Müller, Thomas Deselaers, Eugene Kim, Jayashree Kalpathy-Cramer, Thomas M. Deserno, Paul Clough, and William Hersh. Overview of the ImageCLEFmed 2007 medical retrieval and annotation tasks. In *Working Notes of the 2007 CLEF Workshop*, Budapest, Hungary, September 2007.
6. Henning Müller, Thomas Deselaers, Thomas Lehmann, Paul Clough, Eugene Kim, and William Hersh. Overview of the ImageCLEFmed 2006 medical retrieval and annotation tasks. In *Working Notes of the 2006 CLEF Workshop*, Alicante, Spain, September 2006.