Chapter 12
The Medical Image Classification Task

Tatiana Tommasi and Thomas Deselaers

Abstract We describe the medical image classification task in ImageCLEF 2005–2009. It evolved from a classification task with 57 classes on a total of 10,000 images into a hierarchical classification task with a very large number of potential classes. Here, we describe how the database and the objectives changed over the years and how state–of–the–art approaches from machine learning and computer vision were shown to outperform the nearest neighbor-based classification schemes working on full–image descriptors that were very successful in 2005. In particular the use of discriminative classification methods such as support vector machines and the use of local image descriptors were empirically shown to be important building blocks for medical image classification.

12.1 Introduction

Thanks to the rapid development of modern medical devices and the use of digital systems, more and more medical images are being generated. This has lead to an increase in the demand for automatic methods to index, compare, analyze and annotate them. In large hospitals, several terabytes of new data need to be managed every year. Typically, the databases are accessible only by alphanumeric description and textual meta information through the standard Picture Archiving and Communication System (PACS). This also holds for digital system compliant with the Digital Imaging and Communications in Medicine (DICOM) protocol (Lehmann et al, 2005). The DICOM header contains tags to decode the body part examined, the patient position and the acquisition modality. Some of these are automatically set by

Tatiana Tommasi
Idiap Research Institute, Centre du Parc, Rue Marconi 19, PO Box 592, 1920 Martigny, Switzerland, e-mail: ttommasi@idiap.ch

Thomas Deselaers
Computer Vision Laboratory, ETH Zurich, Zurich, Switzerland,

the digital system according to the imaging protocol used to capture the pixel data. Others are introduced manually by the physicians or radiologists during the routine documentation. This procedure cannot always be considered as reliable, since frequently some entries are either missing, false, or do not describe the anatomic region precisely (Güld et al, 2002). This issue, along with the fact that images may contain semantic information not conveyable by a textual description, has led to growing interest in image data mining and Content–Based Image Retrieval (CBIR). Using information directly extracted from images to categorize them may improve the quality of image annotation in particular, and more generally the quality of patient care.

Until 2005, automatic categorization of medical images was often restricted to a small number of classes. For instance, several algorithms have been proposed for orientation detection of chest radiographs, where lateral and frontal orientation are differentiated by means of digital image processing (Pietka and Huang, 1992; Boone et al, 1992). For this two class experiment, the error rates are below 1% (Lehmann et al, 2003a). Pinhas and Greenspan (2003) report error rates below 1% for automatic categorization of 851 medical images into eight classes. In Keysers et al (2003) six classes are defined according to the body part examined. For their test set of 1,617 images an error rate of 8% is reported. However, such low numbers of classes are not suitable for applications in evidence–based medicine or case–based reasoning. Here the image category must be determined in much more detail.

The ImageCLEF medical image annotation challenge was born in this scenario, proposing a task reflecting real–life constraints of content–based image classification in medical applications. The organizers released a large and heterogeneous x–ray image corpus and invited all the participants to compare their algorithms on it, encouraging advances in the field.

12.2 History of ImageCLEF Medical Annotation

The medical image annotation task was added to the ImageCLEF campaign in 2005 alongside the existing medical retrieval task, and further evolved in its five editions until 2009. A description of the aims and expectations for this task, together with the database used and the error evaluation scheme adopted, is given in the following sections.

12.2.1 The Aim of the Challenge

The aim of automatic image annotation is to describe the image content based on its features, but formally and in a generalized way using methods from pattern recognition and structural analysis. This description can then be used in order to compare a new image to a known data set containing a group of pre–defined classes and thus to assign the correct label to the image.

In the medical area, automatic image classification can help in inserting conventional radiographs into an existing electronic archive without interaction and therefore costly editing of diagnostic findings. Other applications include searching for images in an image database or limiting the number of query results, e.g. after a textual image search. It may even be useful for multi–lingual annotation and DICOM header corrections, or as one component of a diagnosis support system. Without any specific application in mind, the aim of the medical annotation task in ImageCLEF was to evaluate state–of–the–art techniques for automatic annotation of medical images based on their visual properties and to promote these techniques. To provide a fair benchmark, a database of fully classified radiographs was made available to the task participants and could be used to train the classification systems. The challenge consisted of annotating a set of unlabelled images released at a later stage to prevent training on the testing data.

Starting from 2005, the annotation challenge has evolved from a simple classification task with 57 classes to a task with almost 200 classes passing through an intermediate step of about 120 classes. From the very start however, it was clear that the number of classes could not be scaled indefinitely. The number of potential categories that could be recognized in medical applications is far too high to assemble sufficient training data for creating suitable classifiers (Deselaers and Deserno, 2009). One solution to address this issue a hierarchical class structure because it supports the creation of a set of classifiers for subproblems. Therefore, from the very beginning image annotation was based on the hierarchical Image Retrieval in Medical Applications (IRMA) code (see Section 12.2.2).

In 2005 and 2006 the classes were defined by grouping similar codes into single classes and the task was to predict the group to which a test image belongs. In 2007 the objective of the task was refined to predict the complete IRMA code. The hierarchical structure was then used to describe the image content, with the evaluation scheme allowing a finer granularity of classification accuracy. In 2008, high class imbalance was added to promote the function of prior knowledge encoded into the hierarchy. The images in the test set were mainly from classes which had only a few examples in the training data, making annotation significantly harder.

In 2009, for the fifth medical image annotation challenge edition, the task was organized as a survey of the previous year's experience. The idea was to compare the scalability of different image classification techniques with growing numbers of classes, hierarchical class structures and sparsly populated classes.

12.2.2 The Database

The database for the medical image annotation task was provided by the IRMA group from the RWTH University Hospital of Aachen, Germany. It consists of medical radiographs collected randomly from daily routine work at the Department of

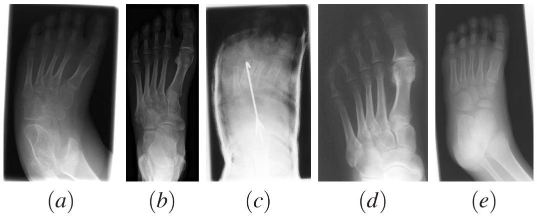


Fig. 12.1: Images from the IRMA database used for the ImageCLEF challenge (Deselaers et al, 2008). Note the high visual variability among the images. They all belong to the same class annotated as: acquisition modality 'overview image'; body orientation 'AP unspecified'; body part 'foot'; biological system 'musculosceletal'.

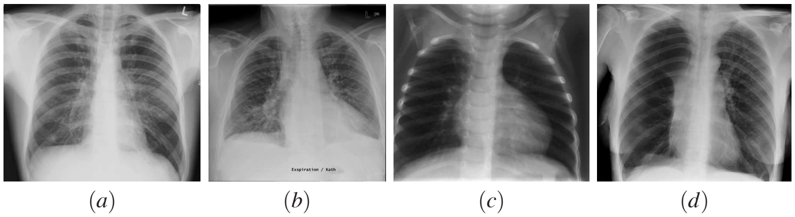


Fig. 12.2: Images from the IRMA database used for the ImageCLEF challenge (Deselaers et al, 2008). Note the high visual similarity between the images. Each of them belongs to a different class. They all have as acquisition modality 'high beam energy', as body region 'chest unspecified', as biological system 'unspecified', but they differ for the body orientation: (a) 'PA unspecified', (b) 'PA expiration' (c) 'AP inspiration', (d) 'AP supine'.

Diagnostic Radiology. Most of the images are secondary digitalized images from plain radiography, but the database also includes images from other modalities, such as CT and ultrasound imaging. The dataset contains a great variability: images of different body parts of patients from different ages, different genders, varying viewing angles, and with or without pathologies. Moreover the quality of radiographs varies considerably and there is a great within–category variability together with a strong visual similarity between many images belonging to different classes (see Figures 12.1 and 12.2). All images were provided as PNG files, scaled to fit into a 512 x 512 pixel bounding box (keeping aspect ratio) using 256 gray values.

In order to establish a ground truth, the images were manually classified by expert physicians using the IRMA code (Lehmann et al, 2003b). This method overcomes the problems of ambiguous and undetailed existing schemes considering 'is a' and 'part of' as the only possible relations between code and sub-code elements. Four aspects of the image acquisition are considered resulting in four axes:

- 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;

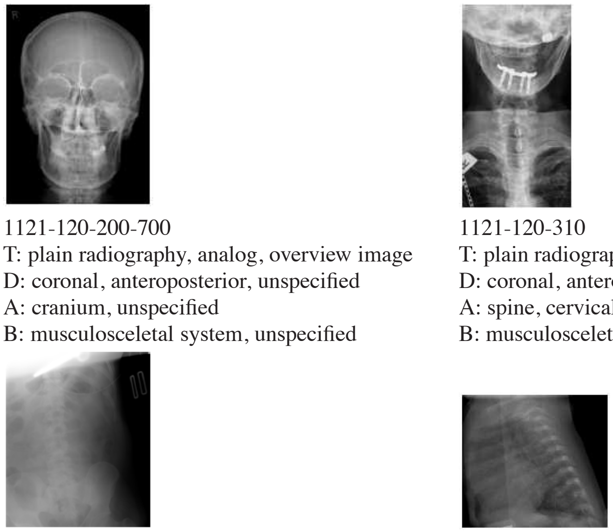
Table 12.1: Examples from the IRMA code, anatomy axis.

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
...	

- the biological code (B) describes the biological system examined.

Each of them is associated with a tag with three to four characters in $\{0, \ldots, 9, \text{a}, \ldots, \text{z}\}$, where '0' denotes 'unspecified' to determine the end of a path along an axis. In this hierarchy, the more the code position differ from '0', the more detailed is the description. Thus the complete IRMA code is a string of 13 characters TTTT-DDD-AAA-BBB, a structure which can be easily extended by introducing characters in a certain code position if new image modalities are introduced. A small excerpt from the anatomy axis of the IRMA code is given in Table 12.1. Exemplar images from the database together with textual labels and their complete code are given in Figure 12.3.

In 2005, a database of 10,000 images was established. To ease the task participation, images were grouped according to their IRMA annotation at a coarse level of detail forming 57 classes. 9,000 randomly chosen images were selected as training data and given to registered participants prior to the evaluation. A remaining set of 1,000 images was published later as test data without category information. Performance was computed on the 1,000 test images and systems compared according to their ability to correctly annotate these images. In all the subsequent edition of the ImageCLEF challenge, the database was built on top of the previous year. In 2006, the 2005 set of 10,000 images was used for training and a new group of 1000 images was collected for testing. The number of classes was more than doubled: based on the IRMA code 116 categories were defined. In 2007, the same procedure was adopted: a new set of 1,000 test images was added and the 11,000 images from 2006 were used as training data. The number of classes remained fixed at 116 but this time the task was not to predict the exact class, but to predict the code and a hierarchy–aware evaluation criterion was defined. In 2008 the data released to participants consisted of 12,076 training images (11,000 training images of 2007 + 1,000 testing images of 2007 + 76 new images) and a new test set of 1,000 samples all annotated with a total of 196 unique codes.



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

1121-120-310
T: plain radiography, analog, overview image
D: coronal, anteroposterior, unspecified
A: spine, cervical spine
B: musculosceletal system, unspecified

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

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

Fig. 12.3: Examples of images and corresponding labels of the IRMA database.

In all the databases used the classes were unevenly distributed reflecting the radiological routine acquisition. However in the first three editions of the challenge, each class contained at least ten images. In 2005, the largest class had 28.6% (2,860 images) share of the complete data set, the second one made up 9.6% (959 images) of the collection and there were several classes that formed only between 0.1% and 0.2% (10 to 20 images) of the complete set (Deselaers et al, 2007). In 2006 the two most populated classes had respectively 19.3% and 9.2% share of the data set, while six classes had only 1% or less (Müller et al, 2006).

Imbalance was worsened in 2008: of the total of 196 codes present in the training stage, only 187 appeared in the test set. 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. The distribution of the test data was nearly uniform while for the training data the distribution was peaked on some classes (Deselaers and Deserno, 2009).

Finally in 2009 a database of 12,677 fully classified radiographs was made available as a training set (Tommasi et al, 2009). Images were provided with labels according to the classification schemes of the annotation tasks from 2005–2008:

- 57 classes as in 2005 (12,631 images) + a ‘clutter’ class C (46 images);
- 116 classes as in 2006 (12,334 images) + a ‘clutter’ class C (343 images);
- 116 IRMA codes as in 2007 (12,334 images) + a ‘clutter’ class C (343 images);

- 193 IRMA codes as in 2008 (12,677 images).

The ‘clutter’ class for a specific setting contained all the images not identifiable in that year but annotated with a higher level of code detail in the subsequent years. The test data consisted of 1,733 images. Not all the training classes have examples in this set:

2005 labels 55 classes (of 57) with 1,639 images + class C with 94 images;
2006 labels 109 classes (of 116) with 1,353 images + class C with 380 images;
2007 labels 109 IRMA codes (of 116) with 1,353 images + class C with 380 images;
2008 labels 169 IRMA codes (of 193) with 1,733 images.

Participating groups were asked to label images according to each of these schemes in order to understand how the hierarchy changes the task and how sparsely populated classes impact performance.

In 2009, the smallest class in the training data contained six images for the 2005–2007 set–ups, and only one image in the 2008 set–up. A total of 20% of the test images belong to sparsely populated training classes. Examples of the different labels are given in Figure 12.4.

12.2.3 Error Evaluation

To evaluate the performance of the runs submitted by the participants to the medical image annotation task, an error evaluation score was defined, which changed in the different editions of the ImageCLEF campaign according to the given image annotations.

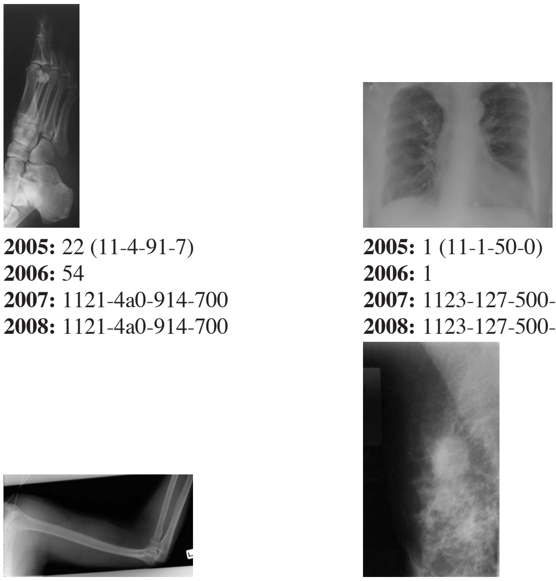
In 2005 and 2006 the error was evaluated just on the capability of the algorithm to make the correct decision. Runs were ranked according to their error rates. For 2007 and 2008, the error was evaluated considering the hierarchical IRMA code.

Let an image be coded by the technical, directional, anatomical and biological independent axes. They can be analyzed separately, summing the error over the individual axes:

- 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 $\hat{l}_i$ all succeeding decisions are considered to be wrong and, given a not–specified position, all succeeding decisions are considered to be not specified. Furthermore, no error is counted if the correct code is unspecified and the predicted code is a wildcard. In that case, all remaining positions are regarded as not specified.



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 in 2009: 50 (11-2-45-7)
2006 in 2009: C
2007 in 2009: CCCC-CCC-CCC-CCC
2008 in 2009: 1121-230-451-700

2005 in 2009: C
2006 in 2009: C
2007 in 2009: CCCC-CCC-CCC-CCC
2008 in 2009: 1127-310-600-625

Fig. 12.4: Examples of all the label settings in the different editions of the medical image annotation task in ImageCLEF.

Wrong decisions that are easy (fewer possible choices at that node) are penalized over wrong decisions that are difficult (many possible choices at that node). 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, wrong decisions at an early stage in the code (higher up in the hierarchy) are penalized more than 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 together:

$$\sum_{i=1}^{I} \underbrace{\frac{1}{b_i}}_{(a)} \underbrace{\frac{1}{i}}_{(b)} \underbrace{\delta(l_i,\hat{l}_i)}_{(c)} \tag{12.1}$$

with

$$\delta(l_i,\hat{l}_i) = \begin{cases} 0 \text{ if } l_j = \hat{l}_j \ \forall j \leq i \\ 0.5 \text{ if } l_j = * \ \exists j \leq i \\ 1 \text{ if } l_j \neq \hat{l}_j \ \exists j \leq i \end{cases} \tag{12.2}$$

where the parts of the equation:

Table 12.2: Error score evaluation for 2007 and 2008 settings. We are considering just the anatomical axis, the correct label is 463.

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

(a) account for difficulty of the decision at position *i* (branching factor);
(b) account 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 (see Table 12.2).

In 2009, the class ‘clutter’ C was introduced. Even if in the test set there were images belonging to this class, their annotation did not influence the error score for the challenge. Moreover, in 2009 the possibility to use wildcards was given even in the 2005 and 2006 settings.

12.3 Approaches to Medical Image Annotation

The ImageCLEF medical image annotation task attracted strong participation from research groups around the world since its first edition. Some of the groups have a background in data mining and retrieval systems while others specialize in object recognition and detection. The sections that follow analyze the methods according to their image representations, classification methods, the use of the hierarchy, and treatment of the unbalanced class distribution.

12.3.1 Image Representation

How to represent the image content is the first problem to face when defining an automatic annotation system. There are different strategies to extract features from images depending on which is considered the most relevant information to capture. As the x–ray images do not contain any color information, edge, shape and global texture features play an important role in this task and were used by several groups (Bo et al, 2005; Deselaers et al, 2005; Liu et al, 2006; Müller et al, 2005). Various methods used the pixel values directly and accounted for possible deformation of the images (Image Distortion Model, IDM) (Güld et al, 2005; Deselaers et al, 2005). Approaches coming from the object recognition field mostly followed the currently widely adopted assumption that an object in images consists of parts that can be modelled independently. Thus these methods considered local features extracted around interest points and used a wide variety of bag-of-features approaches (Marée et al, 2005; Liu et al, 2006; Tommasi et al, 2007; Avni et al, 2008). Generally the ordering of the visual words is not taken into account and only the frequency of the individual visual word is used to form the feature vectors. However, some groups added the spatial information to patches extracted from images (Deselaers et al, 2006; Avni et al, 2008) after observing that radiographs of a certain body part are typically taken in the same spatial arrangement. Another widely adopted strategy consists of combining different local and global descriptors into a unique feature representation (Bo et al, 2005; Liu et al, 2006; Tommasi et al, 2008a).

12.3.2 Classification Methods

Choosing the classification technique means selecting the rules that form the basis of the annotation process. Many different classification strategies were applied and while in the earlier years nearest neighbor-based approaches were most common and most successful (e.g. (Deselaers et al, 2005; Güld et al, 2005)), in 2006 and later, discriminative approaches such as log–linear models (Deselaers et al, 2006), and decision trees (Setia et al, 2008), as well as Support Vector Machines (Setia et al, 2008; Tommasi et al, 2008a; Avni et al, 2008) became more and more common and outperformed the nearest neighbor–based approaches. In many cases known Content–Based Image Retrieval (CBIR) systems are considered: both GIFT (Müller et al, 2005) and FIRE (Deselaers et al, 2005) are used in the same way. The training images are used as the image database and the test images are used to query it. For each query, the training images are ranked according to their similarity and the nearest neighbor decision rule is applied, i.e. the class of the most similar training image is chosen for every test image. Analogous to feature combination, classifier combination has also been a popular way to improve performance (Rahman et al, 2006; Tommasi et al, 2008b; Avni et al, 2008).

12.3.3 Hierarchy

From 2007, when the entire IRMA code was used for labelling, most of the proposed methods tackled the hierarchy considering four different classifiers, one for each axis. The obtained labels were then associated to give the final annotation (Gass et al, 2007; Setia et al, 2008). Other strategies consisted in defining a single classifier able to manage the knowledge encoded in the class hierarchy. Examples are the introduction of weighted distances in k–nearest neighbor classifiers (Springmann and Schuldt, 2007) or weighted splitting rules in decision trees reflecting the hierarchical error score (Setia et al, 2008). Some groups also proposed the combination of axis wise and flat annotation (the ULG group (Deselaers et al, 2008)) or to integrate the output of different classifiers considering majority voting for the characters in each position of the code (Güld and Deserno, 2007). Given the possibility to use wildcards, classifier combination was used to set a '*' when classifiers disagree (Gass et al, 2007; Güld and Deserno, 2007).

12.3.4 Unbalanced Class Distribution

One of the difficulties of the medical image annotation task was the uneven distribution of samples in the training classes. Most of the proposed strategies handled this problem by using wildcards when confidence is low. There have been only few attempts to tackle the class imbalance directly. One of the approaches focused on feature calculation: the number of patches extracted from each image to build the visual word vocabulary was set as inversely proportional to the number of images in its class (Marée et al, 2005). Another approach adapted the classifier using a k-nearest–neighbors (kNN) algorithm with a different *k* value for each class which took into account the frequency of images within the training set (Zhou et al, 2008). The presence of sparsely populated classes in the original training set was also faced by both successively dividing the data into frequency based sub–groups and training a separate SVM for each of them (Unay et al, 2009), and creating virtual examples (Tommasi et al, 2008a).

12.4 Results

In this section we focus on the methods which produced the best results over the five editions of the ImageCLEF medical image annotation task.

A total of 12 research groups participated in 2005 submitting 44 runs. The first fifteen ranked runs are summarized in Table 12.3.

The best results are obtained using the pixel values of the images directly: either by using deformation models on the complete image (scaled to a fixed size) or by using sparsely sampled image patches. Regarding the classification methods, near-

Table 12.3: Resulting error rates for the first 15 runs submitted in 2005 and 2006. (LRPM: low resolution pixel map; BOW: bag–of–words; thumb: thumbnails; Entr.: entropy; relev.: relevance evaluation on the top N retrieved images; HI: histogram intersection kernel; SPM: Spatial Pyramid Matching kernel; RBF: Radial Basis Function kernel; llc, hlc: low and high level cue combination (Tommasi et al, 2008b); oa,oo: one–vs–all and one–vs–one SVM multi–class extension.)

		2005		
Rank	**Group**	**Features**	**Classifier**	**ER(%)**
1	RWTH-i6	thumb. $X \times 32$ IDM	KNN k=1	12.6
2	RWTH-mi	texture JSD + thumb. $X \times 32$ IDM	KNN k=1	13.3
3	RWTH-i6	image patches, BOW	log-linear model	13.9
4	ULG	image patches	boosting	14.1
5	RWTH-mi	texture JSD + thumb. $X \times 32$ IDM	KNN k=1	14.6
6	ULG	image patches	decision trees	14.1
7	GE	texture, 8 grey lev.	GIFT + KNN k=5	20.6
8	Infocomm	texture + LRPM, llc	SVM oa + RBF	20.6
9	GE	texture, 16 grey lev.	GIFT + KNN k=5	20.9
10	Infocomm	texture + LRPM, llc	SVM oa + RBF	20.6
11	Infocomm	texture + LRPM, llc	SVM oa + RBF	20.6
12	GE	texture, 8 grey lev.	GIFT + KNN k=1	21.2
13	GE	texture, 4 grey lev.	GIFT + KNN k=10	21.3
14	MIRACLE	texture	GIFT + relev. N=20	21.4
15	GE	texture, 16 grey lev.	GIFT + KNN k=1	21.7
		2006		
Rank	**Group**	**Features**	**Classifier**	**ER(%)**
1	RWTH-i6	image patches + position, BOW	log-linear model	16.2
2	UFR	local rel. coocc. matr. 1000 p.	SVM oa + HI	16.7
3	RWTH-i6	image patches + position, BOW	SVM oa + HI	16.7
4	CISMeF	local + global texture + PCA	SVM oa + RBF	17.2
5	CISMeF	local + PCA	SVM oa + RBF	17.2
6	MSRA	global, llc	SVM oo + SPM	17.6
7	CISMeF	local + global texture + PCA	SVM oa + RBF	17.9
8	UFR	local rel. coocc. matr. 800 p.	SVM oa + HI	17.9
9	MSRA	image patches, BOW	SVM oo + SPM	18.2
10	CISMeF	local + PCA	SVM oa + RBF	20.2
11	RWTH-i6	thumb. $X \times 32$ IDM	KNN k=1	20.4
12	RWTH-mi	texture JSD + thumb. $X \times 32$ IDM	KNN k=1	21.5
13	RWTH-mi	texture JSD + thumb. $X \times 32$ IDM ’05	KNN k=1	21.7
14	CINDI	local + global, hlc	SVM oo + RBF (+)	24.1
15	CINDI	local + global, hlc	SVM oo + RBF ($\times$)	24.8

est neighbor methods obtain the best results if an appropriate distance function can be defined. Most of the participating methods come from a CBIR context; however, it can be seen that those methods coming from the image classification and recognition domain field achieve good results (ranks 3, 4). The success of the deformation models by the RWTH Aachen University groups might be partly be due to their working with similar data for several years before the competition.

If we compare the winning and the second classified runs, we can see that they differ for the use of texture Tamura features. The RWTH–mi group considered this cue, comparing two images through the Jensen Shannon Divergence. It seems that adding the texture features does not help the classification, but the result may also be due to an unoptimized choice of the cue combining weights.

In 2006, 12 groups took part in the annotation task submitting 28 runs. Looking at the best 15 results (see Table 12.3) the most interesting observation is that the RWTHi6-IDM system that performed best in the previous year's task (error rate: 12.6%) obtained here an error rate of 20.4%. This decrease in performance can be explained by the larger number of classes. All better–ranked approaches use discriminative models (log–linear models or SVMs), which indicates that discriminative approaches can cope better with higher number of classes than the nearest neighbor classifier if the amount of training data is increased by only 10%.

The best–ranked approach in 2006 is a bag–of–visual words model with dense feature extraction and a dense generic visual vocabulary of 65536 visual words incorporating the positions where features were extracted. The position information seems to be very useful and the results validate the hypothesis that as radiographs are taken under controlled conditions, the geometric layout of images showing the same body region can be assumed to be very similar. The second and third ranked approaches also incorporate spatial information into the feature vector.

Considering all the submissions in general, it can be noticed that there is an increasing trend towards the combination of multiple cues at different levels in the classification process.

In 2007, ten groups participated submitting 68 runs. Analyzing the results, it can be observed that the top performing submissions do not consider the hierarchical structure of the given task, but rather use each individual code as a whole and train a 116 class classifier (see Table 12.4). The best run using the code is on rank 6; it builds on top of the other runs from the same group using the hierarchy only in a second stage to put a wildcard where their output differs. Furthermore it can be seen that for a method, which is applied once accounting for the hierarchy/axis structure of the code and once using the straight–forward classification into 116 class approach, the one which does not know about the hierarchy outperforms the other one (runs on ranks 11 and 13, 7 and 14). Another clear observation is that methods using local image descriptors produce better results than 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 global descriptors. The winning run proposes very efficient local features (Scale Invariant Feature Transform (SIFT) (Lowe, 1999) combined with global cues through a new method which performs the integration during the classification process. The method which was ranked best in 2006 was ranked 8 in 2007.

Considering the rank with respect to the applied hierarchical measure, and the ranking with respect to the error rate, it can be seen that they are quite similar. Most of the differences are clearly due to the use of the wildcard characters which can lead to an improvement for the hierarchical evaluation scheme, but will always lead to a deterioration with respect to the the error rate.

In 2008, six groups participated in the image annotation task submitting 24 runs. The 15 fifteen ranked runs use discriminative models and local descriptors outperforming all the other approaches (see Table 12.4). The winning run proposes the combination of local and global features together with a technique to increase the number of images in the scarcely populated classes and evaluates the confidence of the classification decision to use wildcards opportunely. The runs on the sixth and the seventh rank positions respectively by the Idiap and TAU group, use similar features and classification methods and obtain a similar error score. This indicates that the higher performance of the first five runs is most likely due to the use of multiple cues and to the technique adopted to manage the class imbalance and to exploit the hierarchical code structure.

In 2009, seven groups took part to the challenge submitting 19 runs. The task in this last edition was to annotate a set of x–ray images using the labelling schemes from 2005, 2006, 2007 and 2008. Table 12.5 summarizes the best 15 results considering the error score sum on the four different annotation codes. The runs reaching the highest position in the rank are again by the TAU and Idiap groups as in 2008. The results indicate that their strategy suited all the data set configurations proposed in the different editions of the annotation task. In particular the runs submitted by the Idiap group are exactly the same as those in 2008, while the TAU group improved their image descriptors and optimized the kernel choice.

Finally, to evaluate the performances of the best submitted runs all over the five editions of the ImageCLEF annotation task, we can use the RWTH–mi submissions as a reference. This group participated in the competition every year proposing a baseline run (texture JSD + thumb. $X \times 32$ IDM). The ratios between the results of the first ranked run and this baseline submission are reported in Table 12.6. Remember that the error score is defined to be 1 if the code annotation of one image is completely wrong, so the ratios of two error scores can be considered as containing the same information for the ratio between two error rates. The results show an improvement over the years and give clear evidence of the advances obtained in the medical image annotation field.

12.5 Conclusion

The medical image annotation task in ImageCLEF 2005–2009 has established a standard benchmark for medical image annotation. Over the years, the task was developed from a simple classification task into a hierarchical image annotation task with a strongly imbalanced distribution of training images. By comparing the performance of the best ranked run in each year with a baseline method that was submitted in every edition, we have shown that medical image annotation has substantially advanced in the last five years.

Acknowledgements The authors would like to thank Thomas M. Deserno for providing the data for the medical annotation tasks from 2005–2009. We would also like to thank the organizers of CLEF and ImageCLEF for supporting the medical image annotation task over the years. T.T. was supported by the EMMA project thanks to the Hasler foundation (`http://www.haslerstiftung.ch`).

References

Avni U, Goldberger J, Greenspan H (2008) TAU MIPLAB at ImageClef 2008. In: Working Notes of CLEF 2008, Aarhus, Denmark

Bo Q, Wei X, Qi T, Chang SX (2005) Report for annotation task in ImageCLEFmed 2005. In: Working Notes of CLEF 2005, Vienna, Austria

Boone J, Seshagiri S, Steiner R (1992) Recognition of chest radiograph orientation for picture archiving and comunication system display using neural networks. Journal of Digital Imaging 5(3):190–193

Deselaers T, Deserno TM (2009) Medical image annotation in ImageCLEF 2008. In: CLEF 2008 Proceedings, Springer, Lecture Notes in Computer Science (LNCS), vol 5706, pp 523–530

Deselaers T, Weyand T, Keysers D, Macherey W, Ney H (2005) FIRE in ImageCLEF 2005: Combining content–based image retrieval with textual information retrieval. In: CLEF 2005 Proceedings, Springer, Lecture Notes in Computer Science (LNCS), vol 4022, pp 652–661

Deselaers T, Weyand T, Ney H (2006) Image retrieval and annotation using maximum entropy. In: CLEF 2006 Proceedings, Springer, Lecture Notes in Computer Science (LNCS), vol 4730, pp 725–734

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

Deselaers T, Deserno TM, Müller H (2008) Automatic medical image annotation in ImageCLEF 2007: Overview, results, and discussion. Pattern Recognition Letters 29(15):1988–1995

Gass T, Weyand T, Deselaers T, Ney H (2007) FIRE in ImageCLEF 2007: Support vector machines and logistic models to fuse image descriptors for photo retrieval. In: CLEF 2007 Proceedings, Springer, Lecture Notes in Computer Science (LNCS), vol 5152, pp 492–499

Güld MO, Deserno TM (2007) Baseline results for the ImageCLEF 2007 medical automatic annotation task using global image features. In: CLEF 2007 Proceedings, Springer, Lecture Notes in Computer Science (LNCS), vol 5152, pp 637–640

Güld MO, Kohnen M, Keysers D, Schubert H, Wein BB, Bredno J, Lehmann TM (2002) Quality of DICOM header information for image categorization. In: Proceedings SPIE, vol 4685, pp 280–287

Güld MO, Christian Thies BF, Lehmann TM (2005) Combining global features for content–based retrieval of medical images. In: Working Notes of CLEF 2005, Vienna, Austria

Keysers D, Dahmen J, Ney H (2003) Statistical framework for model–based image retrieval in medical applications. Journal of Electronic Imaging 12(1):59–68

Lehmann TM, Güld O, Keysers D, Schubert H, Kohnen M, Wein BB (2003a) Determining the view position of chest radiographs. Journal of Digital Imaging 16(3):280–291

Lehmann TM, Schubert H, Keysers D, Kohnen M, Wein BB (2003b) The IRMA code for unique classification of medical images. In: Proceedings SPIE, vol 5033, pp 440–451

Lehmann TM, Güld MO, Deselaers T, Keysers D, Schubert H, Spitzer K, Ney H, Wein B (2005) Automatic categorization of medical images for content–based retrieval and data mining. Computerized Medical Imaging and Graphics 29(2):143–155

Liu J, Hu Y, Li M, Ma S, ying Ma W (2006) Medical image annotation and retrieval using visual features. In: CLEF 2006 Proceedings, Springer, Lecture Notes in Computer Science (LNCS), vol 4730, pp 678–685

Lowe DG (1999) Object recognition from local scale–invariant features. In: Proceedings of the international conference on computer vision, vol 2, p 1150

Marée R, Geurts P, Piater J, Wehenkel L (2005) Biomedical image classification with random subwindows and decision trees. In: Proceedings of the international conference on computer vision, workshop on Computer Vision for Biomedical Image Applications, Springer, Lecture Notes in Computer Science (LNCS), vol 3765, pp 220–229

Müller H, Geissbühler A, Marty J, Lovis C, Ruch P (2005) The Use of MedGIFT and EasyIR for ImageCLEF 2005. In: CLEF 2005 Proceedings, Springer, Lecture Notes in Computer Science (LNCS), vol 4022, pp 724–732

Müller H, Deselaers T, Deserno T, Kim E, Hersh W (2006) Overview of the ImageCLEFmed 2006 medical retrieval and annotation tasks. In: CLEF 2006 Proceedings, Springer, Lecture Notes in Computer Science (LNCS), vol 4730, pp 595–608

Nilsback M, Caputo B (2004) Cue integration through discriminative accumulation. In: Proceedings of the international conference on computer vision and pattern recognition, vol 2, pp 578–585

Pietka E, Huang H (1992) Orientation correction of chest images. Journal of Digital Imaging 5(3):185–189

Pinhas A, Greenspan H (2003) A continuous and probabilistic framework for medical image representation and categorization. In: Proceedings SPIE, vol 5371, pp 230–238

Rahman MM, Sood V, Desai BC, Bhattacharya P (2006) CINDI at ImageCLEF 2006: Image retrieval and annotation tasks for the general photographic and medical image collections. In: CLEF 2006 Proceedings, Springer, Lecture Notes in Computer Science (LNCS), vol 4730, pp 715–724

Setia L, Teynor A, Halawani A, Burkhardt H (2008) Grayscale medical image annotation using local relational features. Pattern Recognition Letters 29(15):2039–2045

Springmann M, Schuldt H (2007) Speeding up IDM without degradation of retrieval quality. In: Working Notes of CLEF 2007, Budapest, Hungary

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

Tommasi T, Orabona F, Caputo B (2008a) CLEF2008 Image Annotation Task: an SVM Confidence–Based Approach. In: Working Notes of CLEF 2008, Aarhus, Denmark

Tommasi T, Orabona F, Caputo B (2008b) Discriminative cue integration for medical image annotation. Pattern Recognition Letters 29(15):1996–2002

Tommasi T, Caputo B, Welter P, Güld MO, Deserno TM (2009) Overview of the CLEF 2009 medical image annotation track. In: Working Notes of CLEF 2009, Corfu, Greece

Unay D, Soldea O, Ozogur-Akyuz S, Cetin M, Ercil A (2009) Medical image retrieval and automatic annotation: VPA–SABANCI at ImageCLEF 2009. In: Working Notes of CLEF 2009, Corfu, Greece

Zhou X, Gobeill J, Müller H (2008) MedGIFT at ImageCLEF 2008. In: Working Notes of CLEF 2008, Aarhus, Denmark

Table 12.4: Resulting error rates for the first fifteen submitted runs in 2007 and 2008. The two reference run of the RWTH–mi group are also listed. (BOW: bag–of–words; HI: histogram intersection kernel; MCK: multi cue kernel (Tommasi et al, 2008b); DAS: Discriminative Accumulation Scheme (Nilsback and Caputo, 2004); llc, mlc, hlc: low, mid and high level cue combination (Tommasi et al, 2008b); (prob.)= probability interpretation of SVM output; (vote)= combination of voting in one–vs–one multiclass SVM; AX: four different classifications, one for each axis of the IRMA code are performed and then combined; comm.: in the combination of more opinions a wildcard is used where they disagree; virt. imm.: use of virtual samples defined slightly modifying the original images; major: combination on the basis of majority voting; χ^2; chi–square kernel; RBF: Radial Basis Function kernel; Entr: entropy; oa,oo: one–vs–all and one–vs–one SVM multiclass extension; RF relevance feedback.)

2007

Rank	Group	Features	Classifier	Score	ER(%)
1	Idiap	local + global mlc	SVM oa + MCK χ^2	26.8	10.3
2	Idiap	local + global mlc	SVM oo + MCK χ^2	27.5	11.0
3	Idiap	local	SVM oo +χ^2	28.7	11.6
4	Idiap	local	SVM oa +χ^2	29.5	11.5
5	Idiap	local + global hlc	SVM oa +χ^2 DAS	29.9	11.1
6	RWTH-i6	comb. rank 8, 10, 11, 12	log-linear model	30.9	13.2
7	UFR	local rel. coocc. matr. 1000 p.	SVM + HI	31.4	12.1
8	RWTH-i6	patches + position, BOW	log-linear model	33.0	11.9
9	UFR	local rel. coocc. matr. 800 p.	SVM + HI	33.2	13.1
10	RWTH-i6	patches + position, BOW	log-linear model	33.2	12.3
11	RWTH-i6	patches + position, BOW	log-linear model	34.6	12.7
12	RWTH-i6	patches + position, BOW	log-linear model	34.7	12.4
13	RWTH-i6	patches + position, BOW	log-linear model	44.6	17.8
14	UFR	local rel. coocc. matr.	SVM + HI AX	45.5	17.9
15	UFR	local rel. coocc. matr.	decision tree	47.9	16.9
...					
17	RWTH-mi	texture JSD + thumb. $X \times 32$ + IDM	KNN k=5, comm.	51.3	20.0
18	RWTH-mi	texture JSD + thumb. $X \times 32$ + IDM	KNN k=5, major.	52.5	18.0

2008

Rank	Group	Features	Classifier	Score
1	Idiap	local + global, llc + virt. img.	SVM oa +χ^2 comm.	74.9
2	Idiap	local + global, llc + virt. img.	SVM oa +χ^2	83.5
3	Idiap	local + global, llc	SVM oa +χ^2 comm.	83.8
4	Idiap	local + global, mlc + virt. img.	SVM + MCK oa χ^2 comm.	85.9
5	Idiap	local + global, llc	SVM oa +χ^2	93.2
6	Idiap	local	SVM oa +χ^2	100.3
7	TAU	patches whole img. BOW	SVM oo + RBF	105.8
8	TAU	patches mult. res. BOW, hlc	SVM oo + RBF (prob.)	105.9
9	TAU	patches mult. res. BOW, hlc	SVM oo + RBF (vote)	109.4
10	TAU	patches resized img. BOW	SVM oo + RBF	117.2
11	Idiap	local	SVM oa +χ^2	128.58
12	RWTH-mi	texture JSD + thumb. $X \times 32$ IDM	KNN k=5 major.	182.8
13	MIRACLE	local + global	KNN k=3	187.9
14	MIRACLE	local + global	KNN k=2	190.4
15	MIRACLE	local + global	KNN k=2 +RF	190.4

Table 12.5: Resulting error rates for the first fifteen submitted runs in 2009. In the first part of the Table the sum score is reported for each run, in the second part there are the single error scores for each of the used label settings. (BOW: bag–of–words; llc: low level cue combination (Tommasi et al, 2008b); AX: four different classifications, one for each axis of the IRMA code are performed and then combined; comm.: in the combination of more opinions a wildcard is used where they disagree; virt. imm.: use of virtual samples defined slightly modifying the original images; major: combination on the basis of majority voting; χ^2; chi–square kernel; RBF: Radial Basis Function kernel; oa,oo: one–vs–all and one–vs–one SVM multiclass extension; vcad: voting based approach per axis with chopping letter by letter with descending vote.)

		2009		
Rank	**Group**	**Features**	**Classifier**	**Score**
1	TAU	patches mult. res. BOW, llc	SVM oo $+\chi^2$	852.8
2	Idiap	local + global llc + virt. imm.	SVM oa $+\chi^2$ comm.	899.2
3	Idiap	local + global llc	SVM oa $+\chi^2$ comm.	899.4
4	Idiap	local + global llc	SVM oa $+\chi^2$	1039.6
5	Idiap	local + global llc + virt. imm.	SVM oa $+\chi^2$	1042.0
6	FEITIJS	local + global llc	bagging, rand. forest, AX	1352.6
7	VPA-Sabanci	local + block position	SVM oa + RBF, AX	1456.2
8	VPA-Sabanci	local + block position	SVM oa + RBF	1513.9
9	VPA-Sabanci	local + block position freq.	SVM oa + RBF	1554.8
10	VPA-Sabanci	local + block position freq.	SVM oa + RBF	1581.7
11	GE	texture, 8 grey lev. vcad	GIFT + KNN k=5	1633.3
12	GE	texture, 16 grey lev. vcad	GIFT + KNN k=5	1633.3
13	RWTH-mi	texture JSD + thumb. $X \times 32$ IDM	KNN k=5 major.	1994.8
14	GE	texture, 16 grey lev. + SIFT hlc	GIFT + KNN k=5	2097.6
15	VPA-Sabanci	local + block freq.	SVM oa + RBF	2744.1

Rank	Group	2005	2006	2007	2008
1	TAU	356	263	64.3	169.5
2	Idiap	393	260	67.2	178.9
3	Idiap	393	260	67.2	179.2
4	Idiap	447	292	75.8	224.8
5	Idiap	447	292	75.8	227.2
6	FEITIJS	549	433	128.1	242.5
7	VPA-Sabanci	578	462	155.5	261.2
8	VPA-Sabanci	578	462	201.3	272.6
9	VPA-Sabanci	587	498	169.3	300.4
10	VPA-Sabanci	587	502	172.1	320.6
11	GE	618	507	190.7	317.5
12	GE	618	507	190.7	317.5
13	RWTH-mi	790	638	207.6	359.3
14	GE	791.5	612.5	272.7	420.9
15	VPA-Sabanci	587	1170	413.1	574.0

Table 12.6: Resulting error ratios between the best run of each year and the corresponding baseline result. The error ratio for 2009 is evaluated averaging the ratios produced for each of the labelling schemes $(2005, 2006, 2007, 2008)$.

Year	**Group and Run**	**Error Ratio**
2005	RWTH-i6, thumb. $X \times 32$ IDM & FIRE	0.947
2006	RWTH-i6, image patches + position, BOW & FIRE + max Entropy	0.747
2007	Idiap, local + global mlc & SVM oa + MCK χ^2	0.510
2008	Idiap, local + global, llc + virt. img & SVM oa $+\chi^2$ comm.	0.410
2009	TAU, patches mult. res. BOW, llc & SVM oo $+\chi^2$	0.411