The CLEF 2005 Automatic Medical Image Annotation Task

Thomas Deselaers[1], Henning Müller[2], Paul Clough[3],
Hermann Ney[1], and Thomas M. Lehmann[4]

[1]Human Language Technology and Pattern Recognition,
Computer Science Department, RWTH Aachen University, Germany
{deselaers,ney}@cs.rwth-aachen.de
[2]Medical Informatics Service, Geneva University and Hospitals, Geneva Switzerland
henning.mueller@sim.hcuge.ch
[3]Department of Information Studies, Sheffield University, Sheffield, UK
p.d.clough@sheffield.ac.uk
[4]Department of Medical Informatics, Medical Faculty, RWTH Aachen University, Germany
lehmann@computer.org

Abstract

In this paper, the automatic annotation task of the 2005 CLEF cross-language image retrieval campaign (ImageCLEF) is described. This paper focuses on the database used, the task setup, and the plans for further medical image annotation tasks in the context of ImageCLEF. Furthermore, a short summary of the results of 2005 is given. The automatic annotation task was added to ImageCLEF in 2005 and provides the first international evaluation of state-of-the-art methods for completely automatic annotation of medical images based on visual properties.

The aim of this task is to explore and promote the use of automatic annotation techniques to allow for extracting semantic information from little-annotated medical images. A database of 10.000 images was established and annotated by experienced physicians resulting in 57 classes, each with at least 10 images. Detailed analysis is done regarding the (i) image representation, (ii) classification method, and (iii) learning method. Based on the strong participation of the 2005 campain, future benchmarks are planned.

1 Introduction

Evaluating performance is a very important step in the development and investigation of new research methods. In speech recognition, machine translation and information retrieval, large-scale managed evaluation events are a common way to compare the performance of different systems. Examples are the NIST (National Institute of Standards and Technology) machine translation evaluation[1], the TC-STAR evaluation[2], the International Workshop on Spoken Language Translation (IWSLT) Evaluation[3], and the NIST information retrieval evaluation campaign, called TREC[4] (Text REtrieval Conference). In the field of image processing and recognition, evaluation is only recently becoming adopted: Benchathlon[5] is an initiative for evaluating technologies including image filtering, content-based image retrieval (CBIR) and automatic description of images in large-scale image databases. However, to date no evaluation campaign has been carried out. ImageEVAL[6] has done a preliminary test evaluation and is preparing for an official evaluation campaign. TRECVID[7] is an evaluation campaign for video retrieval in the context of the Text REtrieval Conference (TREC)[8] and has been organizing annual benchmarking events since 2001. In the context of the PASCAL network of excellence[9], evaluation campaigns for object classification, detection, and segmentation methods were carried out in March 2005 [1] and in April/May 2006.

The Cross Language Evaluation Forum (CLEF)[10] aims at supporting global digital library applications by developing an infrastructure for testing, tuning, and evaluation of 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 focusses on multi-lingual and more recently on multi-modal aspects of information retrieval. ImageCLEF[11] 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 is searched using textual queries written in a different language. ImageCLEF 2003 attracted just four participants, approaches using a range of text-based retrieval and query enhancement techniques. In 2004, a medical and an interactive retrieval task were added to ImageCLEF. The medical task used a set of images with associated medical case notes and was primarily offered as

[1] `http://www.nist.gov/speech/tests/mt/`

[2] `http://tc-star.org`

[3] `http://www.is.cs.cmu.edu/iwslt2005/evaluation.html`

[4] `http://trec.nist.gov`

[5] `http://www.benchathlon.net/`

[6] `http://www.imageval.org/`

[7] `http://www-nlpir.nist.gov/projects/trecvid/`

[8] `http://trec.nist.gov/`

[9] `http://www.pascal-network.org/`

[10] `http://www.clef-campaign.org`

[11] `http://ir.shef.ac.uk/imageclef/`

a query-by-visual-example (QBVE) retrieval task as search tasks supplied by the organisers contained only images and not text. However, participants could involve text in subsequent retrieval iterations and combine both image processing and text-based retrieval methods. ImageCLEF 2004 attracted strong participation from 18 research groups across the world, demonstrating the need for such an evaluation event. In 2005, the herein described medical automatic annotation task was added to ImageCLEF and again, participation increased. A total of 36 groups registered for ImageCLEF, with 26 groups registering for the automatic annotation task. In the end, 12 groups participated in the annotation task, submitting a total of 41 runs.

Automatic annotation of images in general, in particular of medical images, is a topic of great importance and relevance to the medical community. Currently the most relevant areas are:

1. Automatic Parameter Setting for Image Analysis. The variety of imaging modalities, appearances of different body regions, and the different diagnostic aims require medical image analysis to be specially adapted to the problem. For automatic chains of image processing and analysis, the processing modules must be parameterized accordingly. Thus, the need to classify images emerges to automatically select the necessary image processing steps.

2. Consistency Checks for Meta Data. Medical images are usually stored in the digital imaging and communication in medicine (DICOM) standard that also hosts various other meta data. However, a significant portion of this meta data is wrong, especially when it is generated automatically by the imaging device: Güld et al. [6] reported that an approximate 15-20% of medical images that are recorded using DICOM-compliant modalities have incorrectly specified DICOM tags. In particular, coding of body region is frequently incorrect. Thus, another application of automatical classification of medical images is the validation and correction of its meta data.

3. Generation of Text Queries for Retrieval. In picture archiving and communication systems (PACS), information retrieval is based solely on alphanumerical attributes, i.e. text describing the patient, study, etc. With the increasing importance of images in daily medical routine, effective data management is required. By means of automatic image annotation, a textual description generated completely automatic from image content can be used to improve the query result.

These tasks are strongly related to object recognition, and interestingly the methods from one discipline work surprisingly well for the respective other [7]. Reflecting these applications, the automatical image annotation task in the CLEF 2005 campaign aims at comparing and evaluating different approaches for automatically categorizing images.

2 The IRMA Database

In November 2005, the IRMA database[12] consisted of approximately 17,000 medical radiographs that have been collected arbitrarily from daily routine at the Department of Diagnostic Radiology, RWTH Aachen University, Aachen, Germany[13].

In order to establish a ground truth, the images were manually classified according to a mono-hierarchical, multi-axial coding scheme. More specifically, four axes are used to describe the technique (modality), orientation, body part, and biosystem [8]. Each of these axes allows for specification in three or four levels of detail, and manual reference annotation was performed by skilled radiologists. This annotation process was partly computer-assisted by offering a pre-selection of most likely annotations.

To ease participation, in the first year of automatic image annotation in ImageCLEF not the complete IRMA annotation code was used. Instead, images were grouped (according to their annotation) at a coarser level of detail, forming 57 classes. An example image from each of these 57 classes is depicted in Fig. 1. All images were provided as PNG files, scaled to fit into a 512×512 pixel bounding box (keeping aspect ratio) using 256 gray values.

A subset of 10,000 images was used for ImageCLEF 2005. From this, a set of 9,000 randomly selected images (and category information) was selected as training data and given to registered participants prior to the evaluation. The remaining 1,000 images were later published as test data without category information to prevent *training on the testing data.* Performance was computed on the 1,000 test images, and systems compared according to their ability to correctly annotate these images.

According to radiology routine, the classes are unevenly distributed. For instance, the

[12] `http://irma-project.org`

[13] `http://www.rad.rwth-aachen.de`

largest class (frontal chest radiographs) has a 28.6% (2860 images) share of the complete dataset, the second largest class makes up 9.6% (959 images) of the collection, and there are several classes that form only between 0.1% and 0.2% (10 to 20 images) of the complete set. However, the dataset was designed such that each class consists of at least 10 images. Clearly, some of the classes are visually very similar, e.g. class 7 (plain radiography, coronal, radio carpal joint, musculosceletal system) and class 8 (plain radiography, coronal, hand, musculosceletal system) where the only difference is the shown body part. In these classes, the body parts depicted are very similiar as can be seen from the example images for classes 7 and 8 in Figure 1.

3 Results

While ImageCLEF 2004 already attracted strong participation from 18 research groups across the world, in 2005, a total of 36 groups registered for ImageCLEF, with 26 groups registering for the automatic annotation task. In the end, 12 groups submitted results in the annotation task, leading to a total of 41 runs. The group with the highest number of submissions had seven runs; one group submitted just one run. Table 1 lists participating groups and a short description of the method used. References are given for more detailed description on methods.

As a baseline, the priori probability classifier, i.e. choose always the class with the highest number of observations in the training data, leads to an error rate of 71.1%. This error rate means that 711 of the 1000 images to be classified are misclassified. A more reasonable baseline for optical character recognition (OCR) and medical radiographs was suggested by Keysers et al. [9]. It is provided by a nearest neighbor classifier comparing 32×32 256 gray level thumbnails of the images using the Euclidean distance. On this task, the nearest neighbor Euclidean distance classifier achieves an error rate of 36.8%. The best and worst error rate in the evaluation is 12.6% and 73.3%, respectively (Tab. 1). A combination of various of the submitted classifiers could not improve over the best submission.

Obviously, the classification accuracy depends strongly on the specifict class under consideration. The average classification accuracy over all runs for the different classes ranges from 6.3 % to 90.7 % and there is a tendency that classes with fewer training images are more difficult. For example, images from class 2 (“*plain radiography, coronal, facial cranium,* *musculosceletal system*") were frequently misclassified as class 44 ("*plain radiography, other orientation, facial cranium, musculosceletal system*"): an average of 46% of images from class 2 were classified as class 44.

Classes 7 ("*plain radiography, coronal, radio carpal joint, musculosceletal system*") and 8 ("*plain radiography, coronal, handforearm, musculosceletal system*") are frequently misclassified as class 6("*plain radiography, coronal, hand, musculosceletal system*"), where again class 6 is much better represented in the training data. Furthermore, many classes (6, 13, 14, 27, 28, 34, 44, 51, 57) are often misclassified to be from class 12, which is by far the largest class in the training data. This strongly coincides with the fact that class 12 ("*plain radiography, coronal, chest, unspecified*") is the class with the highest classification accuracy: on average 90.7% of the test images from class 12 were classified correctly. The three classes with the lowest classification accuracies, form together less then 1% of the training data. In Figure 2, the average confusion matrix over all submissions is visualized. Here, darker fields denote higher values. As the main diagonal of the matrix has much higher values than the other fields, the classifiers perform well on average

Three criterions are used to analyze the methods:

- **Image Representation.** Several methods directly use the pixel values of the images and account for possible deformations in the images (i.e. ranks 1, 2, and 5). The methods coming from the object recognition field follow the currently widely adopted assumption that objects in images consist of parts that can be modelled independently. Thus, these methods use local features extracted around interest points (i.e. ranks 3, 4, and 5). Other methods use quantization to different numbers of gray levels in combination with Gabor filters (ranks 7, 9, 12, 13, 15, 19, and 20) from the medGIFT[14] image retrieval system. As the images do not contain any color information, texture features like the Tamura texture features [22], the MPEG-7 visual descriptors [23], or Gabor features [24] play an important role for this task and were used by several groups (e.g. ranks 2, 5, 7, 8, 9, 10, ...).

- **Classification Method.** Many of the submitted results were created using k-nearest neighbor classification with k between 1 and 20 (ranks 1, 2, 5, 7, 17, ...). Several methods use the GIFT retrieval metric for the determination of nearest neighbors(ranks

[14] `http://www.sim.hcuge.ch/medgift/`

7,9,12,13,15,19,20,...). Variations like nearest prototype classification and majority voting were also applied. One method uses a maximum entropy classifier (rank 3), one method uses boosting and decision trees (ranks 4 and 6), and some groups use support vector machines (ranks 8, 10, 11, 16, 23, 24, 25, 26, 27, 32, and 36). For support vector machines, the variety in performance is high and probably depends on the image representations and kernels used.

- **Learning Technique.** Explicit learning and training methods were used only by the groups using support vector machines, boosting and decision trees, or the maximum entropy classifier. The other groups used the parameter-free nearest neighbor classifier. Nonetheless, the medGIFT group is planning to use a training method in upcoming annotation tasks [25].

In summary, it can be seen that the best results are obtained using the pixel values of the images directly: either by using sparsely sampled image patches or by using the complete image unchanged. For classification, discriminative approaches seem to perform very well and nearest neighbor classification is also well suited if a suitable distance function can be defined.

In Figure 3 some test images that were correctly classified by all submissions and some test images that were misclassified most frequently are depicted together with their classes.

Discussion. Due to the high participation and the good results that were achieved by several methods, the automatic annotation task in ImageCLEF 2005 appears to have been a great success. The task can be considered to be realistic, as the images have been taken randomly from clinical routine and the problem of correcting annotations of primarily digital images and annotation of secondary digital images is a problem of daily routine. In summary, the database is a valuable resource for testing and creating automatic image annotation systems, and the high participation in ImageCLEF has shown the need for such evaluation.

Although most of the participating methods come from a CBIR context (e.g. the methods ranked 1, 2, 5, 7, 9, 12, 13, 14, ...) it can be seen that those methods that come from the image classification and recognition (e.g. the methods ranked 3, 4, 6, 8, 10, 11) field can achieve excellent results for the task of automatic annotation of medical images. The success of the two groups from RWTH Aachen University might partly be due to their working with similar data for some time before.

Methods using the pixel values directly and deformation models outperform most other methods for the given task. Methods from object recognition, assuming that objects in images can be modelled as a set of parts, also perform very well although they were not tuned with respect to this task. It can also be seen that image retrieval methods perform well for this task, especially if domain knowledge from medicine can be incorporated (ranks 2, 5, 7, ...).

Interestingly, the classifer used is not of such great importance, because the classifiers that were applied are spread over the whole range of submissions. The methods from object recognition have the advantage that training of these methods is a well-investigated area and usually discriminative methods are applied. In contrast, in the CBIR domain training of parameters is still uncommon.

4 Summary and Conclusion

We presented the outcomes of the medical automatic image annotation task of ImageCLEF 2005 and shortly described the methods of the participating groups. The data of the evaluation is available for free to encourage comparison of new approaches to the outcomes of the challenge.

Outlook With ImageCLEFmed 2005, a valuable resource for benchmarking of automatic annotation algorithms has been created. In 2006, this task was extended and continued with a similar number of participants. 10,000 training images from 117 classes were provided and a new set of 1,000 test images. Furthermore, a non-medical automatic annotation task was established in cooperation with the MUSCLE[15] network of excellence[16].

For 2007, it is planned to create a *hierarchical* classification task. That is, the training images are published with their entire IRMA code and a new set of test images have to be classified. Instead of absolute decisions for a class, the classifier will be able to decide on its own to what level of detail the classification is done on which of the annotation axes.

[15] Multimedia Understanding through Semantics, Computation and Learning

[16] `http://www.muscle-noe.org/`

Acknowledgment

This work was partially funded by the Deutsche Forschungsgemeinschaft (DFG) under grands NE-572/6 and Le 1108/4 and by the Swiss National Science Foundation (NSF) grant ITR-0325160 and grant 205321-109304. Furthermore we would like to thank the CLEF campaign for allowing ImageCLEF to be part of it.

References

[1] Everingham M, Zisserman A, Williams CKI, van Gool L, Allan M, et al. The 2005 PASCAL visual object classes challenge. In: Selected Proceedings of the first PASCAL Challenges Workshop. Lecture Notes in Artificial Intelligence (to appear). Southampton, UK: Springer; 2006.

[2] Müller H, Michoux N, Bandon D, Geissbuhler A. A review of content-based image retrieval systems in medical applications – clinical benefits and future directions. International Journal of Medical Informatics 2004;73:1–23.

[3] Tourassi GD. Journey toward computer-aided diagnosis: role of image texture analysis. Radiology 1999;(213):317–320.

[4] Chen DR, Chang RF, Huang YL. Computer-aided diagnosis applied to us of solid breast nodules by using neural networks. Radiology 1999;213:407–412.

[5] Yamamoto S, Jiang H, Matsumoto M, Tateno Y, Iinuma T, Matsumoto T. Image processing for computer-aided diagnosis of lung cancer by CT. In: 3rd IEEE Workshop on Applications of Computer Vision. Sarasota, FL, USA; 1996. p. 236–241.

[6] Güld MO, Kohnen M, Keysers D, Schubert H, Bredno J, Lehmann TM. Quality of DICOM header information for image categorization. In: SPIE 2002. vol. 4685; 2002. p. 280–287.

[7] Clough P, Mueller H, Deselaers T, Grubinger M, Lehmann TM, Jensen J, et al. The CLEF 2005 cross-language image retrieval track. In: Workshop of the Cross–Language Evaluation Forum (CLEF 2005). LNCS. Vienna, Austria; 2005. to appear in 2006.

[8] Lehmann TM, Schubert H, Keysers D, Kohnen M, Wein BB. The IRMA code for unique classification of medical images. In: Proc. SPIE 2003. vol. 5033; 2003. p. 440–451.

[9] Keysers D, Dahmen J, Ney H, Wein B, Lehmann TM. Statistical framework for model-based image retrieval in medical applications. Journal of Electronic Imaging 2003;12(1):59–68.

[10] Keysers D, Gollan C, Ney H. Classification of medical images using non-linear distortion models. In: Proc. BVM 2004, Bildverarbeitung für die Medizin. Berlin, Germany; 2004; 366–370.

[11] Lehmann TM, Güld MO, Deselaers T, Keysers D, Schubert H, Spitzer K, Ney H, Wein BB Automatic categorization of medical images for content-based retrieval and data mining. Computerized Medical Imaging and Graphics 2005;29:143–155.

[12] Deselaers T, Keysers D, Ney H. Discriminative training for object recognition using image patches. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 05). San Diego, CA; 2005; 2:157–162.

[13] Marée R, Geurts P, Piater J, Wehenkel L. Biomedical image classification with random subwindows and decision trees. In: Proc. ICCV workshop on Computer Vision for Biomedical Image Applications (CVIBA 2005). Lecture Notes in Computer Science 3765; 2005. p. 220–229.

[14] Müller H, Geissbuhler A, Marty J, Lovis C, Ruch P. The use of medGIFT and easyIR for ImageCLEF 2005. In: Proceedings of the Cross Language Evaluation Forum 2005. Lecture Notes in Computer Science. Vienna, Austria; 2006; to appear.

[15] Qiu B, Xiong W, Tian Q, Xu CS. Report on the annotation task in ImageCLEFmed 2005. In: Working Notes of the CLEF Workshop 2005. Vienna, Austria; 2005. .

[16] Villena-Román J, González-Cristóbal JC, Goñi-Menoyo JM, Martínez-Fernandez JL. MIRACLE's naive approach to medical images annotation. In: Working Notes of the CLEF Workshop 2005. Vienna, Austria; 2005.

[17] Chang YC, Lin WC, Chen HH. Combining text and image queries at ImageCLEF2005. In: Working Notes of the CLEF Workshop 2005. Vienna, Austria; 2005.

[18] Cheng PC, Chien BC, Ke HR, Yang WP. NCTU-DBLAB at ImageCLEF 2005: automatic annotation task. In: Working Notes of the CLEF Workshop 2005. Vienna, Austria; 2005.

[19] Besançon R, Millet C. Merging results from different media: Lic2m experiments at ImageCLEF 2005. In: Working Notes of the CLEF Workshop 2005. Vienna, Austria; 2005.

[20] Petkova D, Ballesteros L. Categorizing and annotating medical images by retrieving terms relevant to visual features. In: Working Notes of the CLEF Workshop 2005. Vienna, Austria; 2005.

[21] Rahman M, Desai BC, Bhattacharya P. Supervised machine learning based medical image annotation and retrieval. In: Working Notes of the CLEF Workshop 2005. Vienna, Austria; 2005.

[22] Tamura H, Mori S, Yamawaki T. Textural features corresponding to visual perception. IEEE Transaction on Systems, Man, and Cybernetics 1978;8:6:460–472.

[23] Eidenberger H. How good are the visual MPEG-7 features? Proceedings SPIE Visual Communications and Image Processing Conference 2003; Lugano, Italy, 2003, 5150:476–488.

[24] Gabor D. Theory of Communication. Journal of IEE 1946;93:429–457.

[25] Müller H, Squire DM, Pun T. Learning from user behavious in image retrieval: applications of the market basket analysis. International Journal of Computer Vision 2004;56(1-2 (Special Issue on Content-based Visual Information Retrieval):65–77.

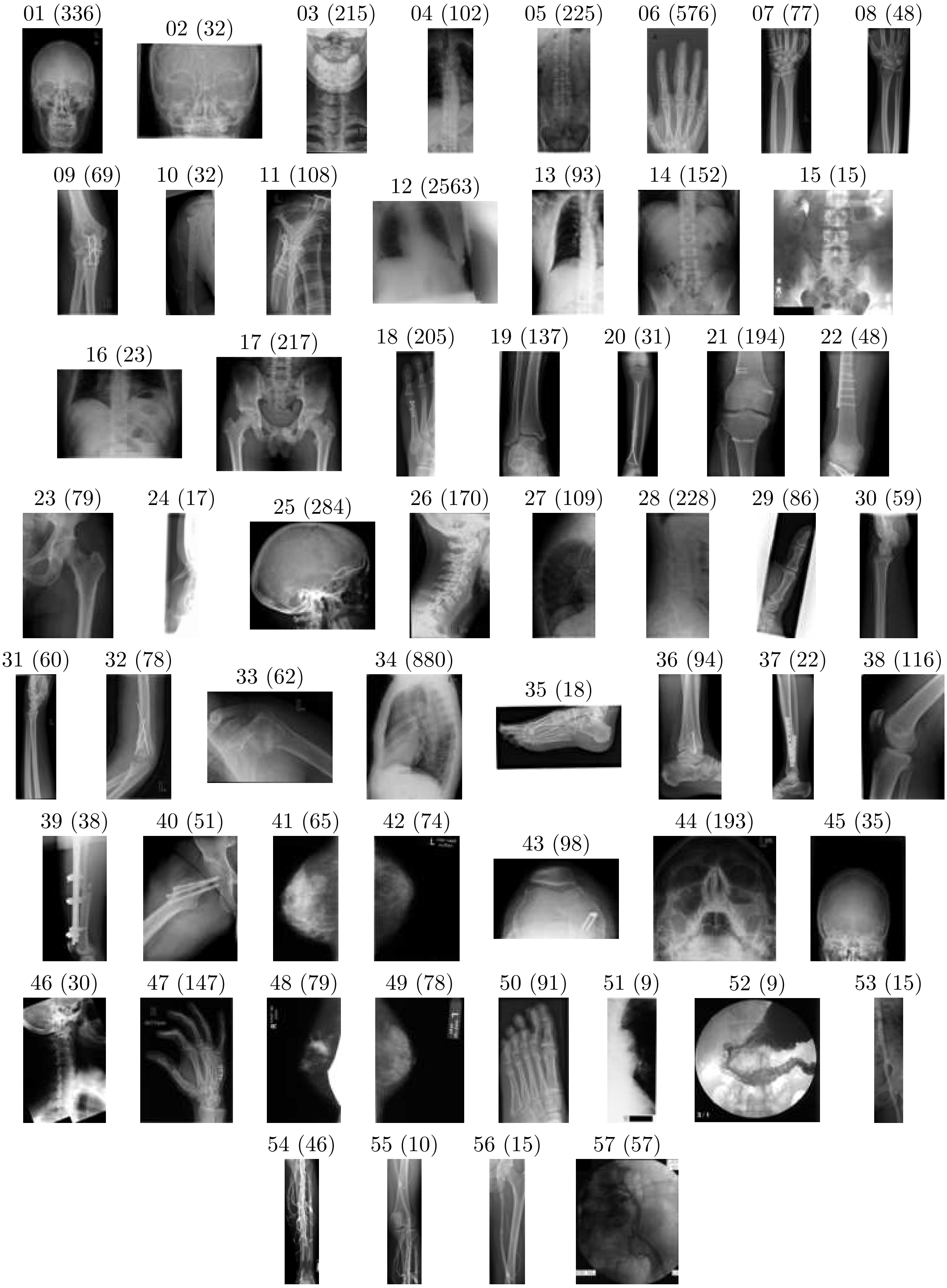

Figure 1: One example from each of the 57 classes (with the number of training examples from this class).

Table 1: Resulting error rates for the submitted runs. (Abbreviations for groups: *U*: University, *CS*: computer science, *UPM*: Universidad Politecnica de Madrid, *DBLAB*: database lab, *NCTU*: National Chiao Tung University, *NTU*: National Taiwan University, *CEA*: Commissariat à l'énergie atomique. Abbreviations for methods: *IDM*: image distortion model, *SVM*: support vector machines, *MedGIFT*: medical GNU image finding tool, *GIFT*: GNU image finding tool, *k-NN*: *k* nearest neighbor.)

rank	group	method	ref.	ER[%]
1	RWTH Aachen U, CS Dep., DE	IDM	[10]	12.6
2	RWTH Aachen U, Med. Inf., DE	IDM & Texture feature	[11]	13.3
3	RWTH Aachen U, CS Dep., DE	image patches & discriminative training	[12]	13.9
4	U Liège, BE	image patches & boosting	[13]	14.1
5	RWTH Aachen U, Med. Inf., DE	IDM & Texture feature	[11]	14.6
6	U Liège, BE	image patches & decision trees	[13]	14.7
7	U & Hospital Geneva, CH	MedGIFT	[14]	20.6
8	Infocom, Singapore, SG	SVM & various image features	[15]	20.6
9	U & Hospital Geneva,CH	MedGIFT	[14]	20.9
10	Infocom, Singapore, SG	SVM & various image features	[15]	20.9
11	Infocom, Singapore, SG	SVM & various image features	[15]	21.0
12	U & Hospital Geneva, CH	MedGIFT	[14]	21.2
13	U & Hospital Geneva, CH	MedGIFT	[14]	21.3
14	Miracle from UPM Madrid, ES	GIFT & majority voting	[16]	21.4
15	U & Hospital Geneva, CH	MedGIFT	[14]	21.7
16	Infocom, Singapore, SG	SVM & various image features	[15]	21.7
17	National Taiwan U, TW	block features & nearest neighbor	[17]	21.7
18	National Taiwan U, TW	block features & top 2 classifier	[17]	21.7
19	U & Hospital Geneva, TW	MedGIFT	[14]	21.8
20	U & Hospital Geneva, TW	MedGIFT	[14]	22.1
21	Miracle from UPM Madrid, ES	GIFT & majority voting	[16]	22.3
22	National Taiwan U, TW	block features & nearest prototype	[17]	22.5
23	DBLAB from NCTU, TW	SVM, various img. feat.	[18]	24.7
24	DBLAB from NCTU, TW	SVM, various img. feat.	[18]	24.9
25	DBLAB from NCTU, TW	SVM, various img. feat.	[18]	28.5
26	DBLAB from NCTU, TW	SVM, various img. feat.	[18]	31.8
27	DBLAB from NCTU, TW	SVM, various img. feat.	[18]	33.8
28	CEA, FR	*k*-NN classifier & image projection	[19]	36.9
29	Mt. Holyoke College, MA, USA	Gabor Energy	[20]	37.8
30	Mt. Holyoke College, MA, USA	Gabor Energy	[20]	40.3
31	CEA, FR	*k*-NN & local edge patterns	[19]	42.5
32	CINDI from Concordia U, CA	SVM, various image feat.	[21]	43.3
33	CEA, FR	*k*-NN & quantified colors	[19]	46.0
34	U Montreal, CA	feature combination		55.7
35	U Montreal, CA	texture coarseness		60.3
36	DBLAB from NCTU, TW	SVM, various img. feat.	[18]	61.5
37	U Montreal, CA	contour image features		66.6
38	U Montreal, CA	shape image features		67.0
39	U Montreal, CA	centered contours		67.3
40	U Montreal, CA	Fourier shape feat.		67.4
41	U Montreal, CA	directionality		73.3
Euclidean Distance, 32x32 images, 1-Nearest-Neighbor				36.8
apriori probability classifier				71.1

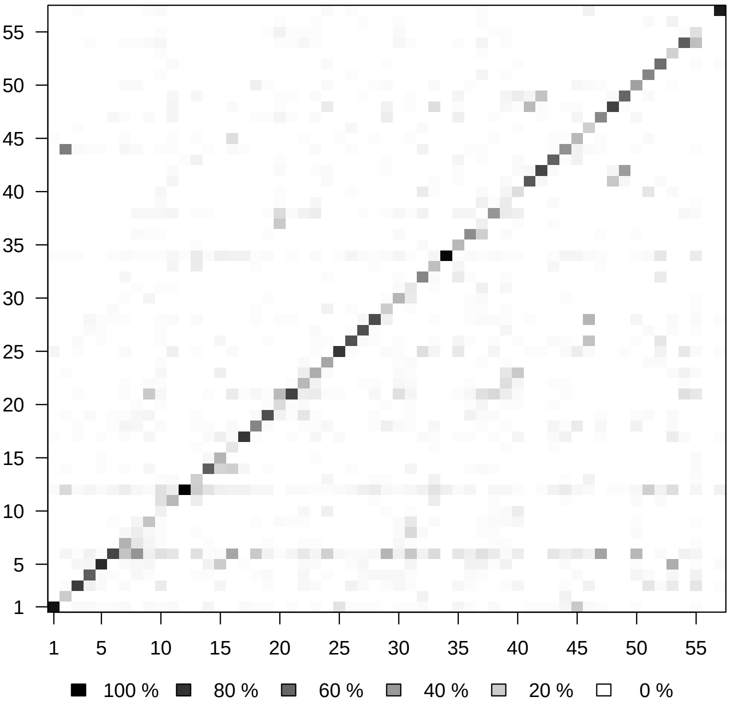

Figure 2: Average confusion matrix over all submitted runs. Dark fields denote high values.

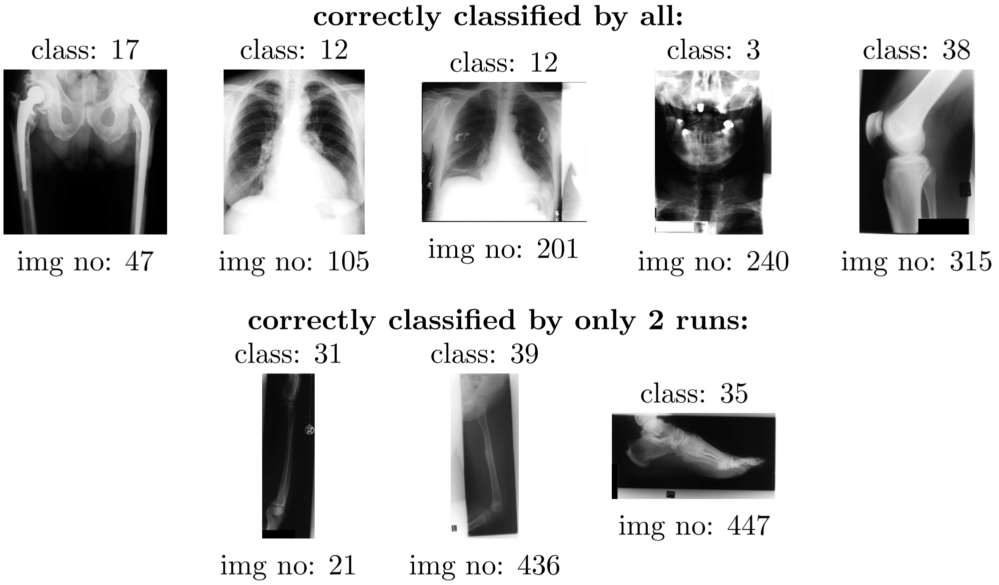

Figure 3: The images that were classified correctly by all systems and the images that were misclassified most often.