Overview of the ImageCLEFmed 2006 Medical Retrieval and Medical Annotation Tasks

Henning Müller[1], Thomas Deselaers[2], Thomas Deserno[3], Paul Clough[4], Eugene Kim[5], and William Hersh[1]

[1] Medical Informatics, University and Hospitals of Geneva, Switzerland
[2] Computer Science Dep., RWTH Aachen University, Germany
[3] Medical Informatics, RWTH Aachen University, Germany
[4] Sheffield University, Sheffield, UK
[5] Oregon Health and Science University (OHSU), Portland, OR, USA
henning.mueller@sim.hcuge.ch

Abstract. This paper describes the medical image retrieval and annotation tasks of ImageCLEF 2006. Both tasks are described with respect to goals, databases, topics, results, and techniques. The ImageCLEFmed retrieval task had 12 participating groups (100 runs). Most runs were automatic, with only a few manual or interactive. Purely textual runs were in the majority compared to purely visual runs but most were mixed, using visual and textual information. None of the manual or interactive techniques were significantly better than automatic runs. The best–performing systems used visual and textual techniques combined, but combinations of visual and textual features often did not improve performance. Purely visual systems only performed well on visual topics. The medical automatic annotation used a larger database of 10,000 training images from 116 classes, up from 9,000 images from 57 classes in 2005. Twelve groups submitted 28 runs. Despite the larger number of classes, results were almost as good as in 2005 which demonstrates a clear improvement in performance. The best system of 2005 would have received a position in the middle in 2006.

Keywords: image retrieval, automatic image annotation, medical information retrieval.

1 Introduction

ImageCLEF[1] [1] started within CLEF (Cross Language Evaluation Forum) in 2003. A medical image retrieval task was added in 2004 to explore domain–specific retrieval as well as multi-modal retrieval (combining visual and textual features for retrieval). Since 2005, a medical retrieval and a medical image annotation task have been parts of ImageCLEF. This paper concentrates on the two medical tasks, whereas a second paper [2] describes the new object classification and the photographic retrieval tasks. More detailed information can also be

[1] http://ir.shef.ac.uk/imageclef/

C. Peters et al. (Eds.): CLEF 2006, LNCS 4730, pp. 595–608, 2007.

found on the task web pages for ImageCLEFmed[2] and the medical annotation task[3]. Detailed analyses of the 2005 medical image retrieval task and of the 2005 medical annotation task are available in [3] and [4], respectively.

2 The Medical Image Retrieval Task

2.1 General Overview

In 2006, the medical retrieval task was run for the third year, and for the second year with the same dataset of over 50,000 images from four collections. One of the most interesting findings for 2005 was the variable performance of systems based on whether the topics had been classified as amenable to visual, textual, or mixed retrieval methods. For this reason, we developed 30 topics for 2006, with 10 each in the three categories. The scope of the topic development was slightly enlarged by using the log files of a medical media search engine of the Health on the Net (HON) foundation. Analysis of these logs showed a great number of general topics not covering the entire four axes defined in 2005:

- Anatomic region shown in the image;
- Image modality (e.g. x–ray, CT, MRI, gross pathology, etc.);
- Pathology or disease shown in the image;
- Abnormal visual observation (e.g. enlarged heart).

The process of relevance judgements was similar to 2005 and trec_eval was used for the evaluation of the results.

2.2 Registration and Participation

In 2006, a record number of 47 groups registered for ImageCLEF and among these, 37 registered for the medical retrieval task. Groups came from four continents and a total of 16 countries. Unfortunately, some registered group did not send in results. 12 groups from 8 countries submitted results. Each entry below describes briefly the techniques used for their submissions.

- *CINDI, Canada.* The CINDI group from Concordia University, Canada, submitted a total of four runs: one purely textual, one visual, and two combined runs. Text retrieval was based on Apache Lucene. For visual information a combination of global and local features were used and compared using Euclidean distance. Most submissions used relevance feedback (RF).
- *MSR, China.* Microsoft Research China submitted one purely visual run using a combination of various features accounting for color and texture.

[2] http://ir.ohsu.edu/image/

[3] http://www-i6.informatik.rwth-aachen.de/~deselaers/imageclef06/medicalaat.html

- *Institute for Infocomm Research I2R–IPAL, Singapore.* IPAL submitted 26 runs, the largest number of any group. Textual and visual runs were prepared in cooperation with I2R. For visual retrieval, patches of image regions were applied and manually classified into semantically valid categories and mapped to Unified Medical Language System (UMLS). For the textual analysis, all query languages were separately mapped to UMLS and then applied to retrieval. Several classifiers based on SVMs and other classical approaches were used and combined.
- *UKLFR, Germany.* The University Hospitals of Freiburg submitted 9 runs mainly using textual retrieval. Interlingua and the original language were used (morphosaurus and Lucene). Queries were preprocessed by removing the "show me" test. Runs differed in language and combination with GIFT.
- *SINAI, Spain.* Jaen University submitted 12 runs: 3 of them using only textual information and 9 using a text retrieval system and adding provided data from the GIFT image retrieval system. The runs differed in settings for "information gain" and the weighting of textual and visual information.
- *OHSU, USA.* Oregon Health and Science University performed manual modification of queries and fusion with results from visual runs. One run established a baseline using the text of the topics as given. Another run then manually modified the topic text removing common words and adding synonyms. For both runs, there were submissions in each of the three individual languages (English, French, German) plus a merged run with all and a run with the English topics expanded with automatic translation using the Babelfish translator. The manual modification of the queries improved performance substantially. The best results came from the English–only queries, followed by the automatically translated and the merged queries. One additional run assessed fusing data from a visual run with the merged queries. This decreased MAP but did improve precision at 10 and 30 images.
- *I2R Medical Analysis Lab, Singapore.* Their submission was together with the IPAL group from the same lab.
- *MedGIFT, Switzerland.* The University and Hospitals of Geneva relied on two retrieval systems for their submission. The visual part was performed with medGIFT. The textual retrieval used a mapping of the query and document text towards concepts in MeSH (Medical Subject Headings). Then, matching was performed with a frequency–based weighting method using easyIR. All results were automatic runs using visual, textual and mixed features. Separate runs were submitted for the three languages.
- *RWTH Aachen University – Computer Science, Germany.* The RWTH Aachen University, CS department, submitted 9 runs, all using the FIRE system and various features describing color, texture, and global appearance. For one run, the queries and the qrels of last year were used as training data to obtain weights for the combination of features using maximum entropy training. One run was purely textual, 3 were purely visual, and the remaining 5 runs used textual and visual information. All runs were fully–automatic.
- *RWTHmi, Germany.* The medical Informatics group at RWTH Aachen submitted 2 purely visual runs without interaction. Both runs used a combination

of global appearance and texture features compared with invariant distance measures. Runs differed in the weights for the features.
- *SUNY, USA.* State University New York submitted 2 purely textual runs and 2 using text and visual information. Parameters for the system were tuned using 2005 topics and automatic RF in variations.
- *LITIS Lab, INSA Rouen, France.* The INSA group from Rouen submitted one run using visual and textual information. MeSH dictionaries were used for text analysis, and the images were represented by various features accounting for global and local information. Most of the topics were treated fully automatic and four topics were treated with manual interaction.

2.3 Databases

In 2006, the same dataset was used as in 2005 containing four sets of images. Casimage was made available, containing almost 9,000 images of 2,000 cases [5]. Casimage includes mostly radiology, but also photographs, PowerPoint slides, and illustrations. Cases are mainly in French, with around 20% being in English and 5% without annotation. We also used PEIR[4] (Pathology Education Instructional Resource) with annotations based on HEAL[5] (Health Education Assets Library, mainly Pathology images [6]). This dataset contains over 33,000 images with English annotations, annotation being per image. The nuclear medicine database of MIR, the Mallinkrodt Institute of Radiology [7], was also distributed containing over 2,000 images mainly from nuclear medicine with annotations provided per case and in English. Finally, PathoPic [8] was included in our dataset containing 9,000 images with extensive annotation per image in German. Part of the German annotation is translated into English. As such, we were able to use a total of more than 50,000 images, with annotations in three different languages. Through an agreement with the copyright holders, we were able to distribute these images to the participating research groups.

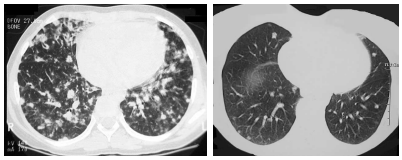
2.4 Query Topics

The query topics were based on two surveys performed in Portland and Geneva [9,10]. In addition to this, a log file of a media search engine HON[6] was used to create topics along the following axes:

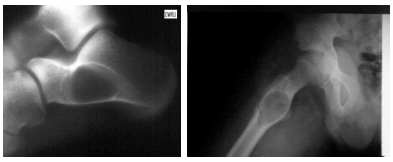
- Anatomic region shown in the image;
- Image modality (x–ray, CT, MRI, gross pathology, etc.);
- Pathology or disease shown in the image;
- Abnormal visual observation (e.g. enlarged heart).

The HON log-files indicated rather general topics than the specific ones used in 2005, so we used real queries from the log-files in 2006. We could not use the most frequent queries, since they were too general, e.g. heart, lung, etc.,

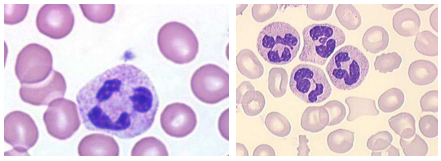
[4] http://peir.path.uab.edu/
[5] http://www.healcentral.com/
[6] http://www.hon.ch/



Show me chest CT images with nodules.
Zeige mir CT Bilder der Lunge mit Knötchen.
Montre–moi des CTs du thorax avec nodules.



Show me x–ray images of bone cysts.
Zeige mir Röntgenbilder von Knochenzysten.
Montre–moi des radiographies de kystes d'os.



Show me blood smears that include polymorphonuclear neutrophils.
Zeige mir Blutabstriche mit polymorphonuklearer Neutrophils.
Montre–moi des échantillons de sang incluant des neutrophiles polymorphonucléaires.

Fig. 1. Examples for a visual, a mixed and a semantic topic

but those that satisfied at least two of the axes. After identifying 50 candidate topics, we grouped them into three classes based upon an estimation of what retrieval techniques they would be most retrievable – visual, mixed, or textual. Another goal was to cover frequent diseases and have a balanced variety of imaging modalities and anatomic regions. After choosing 10 queries for each category, we manually searched query images on the web. In 2005, images were taken partly from the collection. Although they were most often cropped, having external images made the visual task more challenging, as these images could be from other modalities and have completely different characteristics. Figure 1 shows examples for visual, mixed and semantic topics.

2.5 Relevance Judgements

For relevance judging, pools were built from all images for a given topic ranked in the top 30 retrieved. This gave pools from 647 to 1,187 images, with a mean of 910 per topic. Relevance judgements were performed by seven US physicians enrolled in the OHSU biomedical informatics graduate program. Eleven of the 30 topics were judged in duplicate, with two judged by three different judges. Each topic had a designated "original" judge from the seven. A total of 27,306 relevance judgements were made. (These were primary judgements; ten topics had duplicate judgements.) The judgements were turned into a qrels file, which was then used to calculate results with trec_eval. We used Mean Average Precision (MAP) as the primary evaluation measure. We note, however, that its orientation to recall (over precision) may may not be appropriate for many image retrieval tasks.

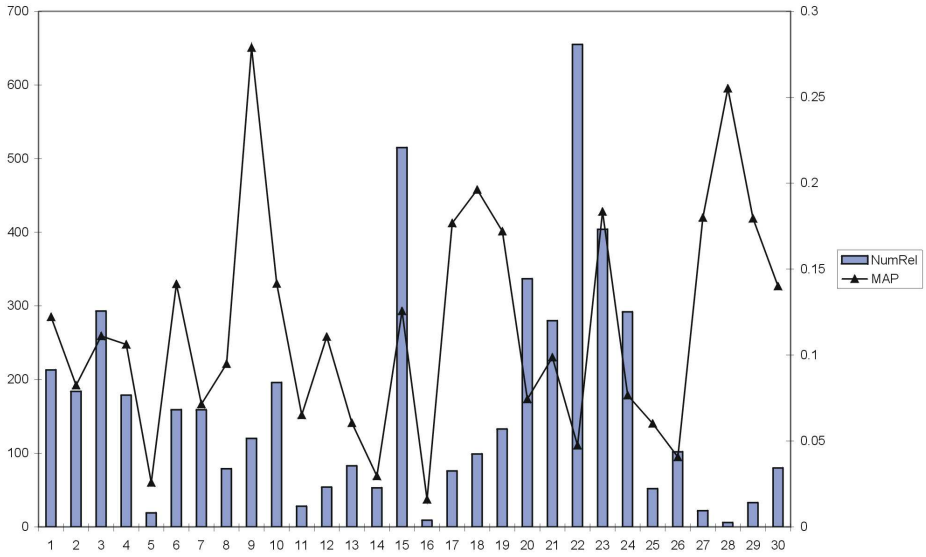


Fig. 2. Evaluation results and number of relevant images per topic

2.6 Submissions and Results

12 groups from eight countries participated in ImageCLEFmed 2006. These groups submitted 100 runs, with each group submitting from 1 to 26 runs.

We defined two categories for the submitted runs: one for the interaction used (automatic – no human intervention, manual – human modification of the query before the output of the system is seen, and interactive – human modification of the query after the output of the system is seen) and one for the data used for retrieval (visual, textual, or a mixture). The majority of submitted runs were automatic. There were fewer visual runs than there were textual and mixed runs.

Figure 2 gives an overview of the number of relevant images per topic and of the performance that this topic obtained on average (MAP). It can be seen that the variation in this case was substantial. Some topics had several hundred relevant images in the collection, whereas others only had very few. Likewise, performance could be extremely good for a few topics and extremely bad for others.

Figure 3 illustrates a comparison of several measurements for all submitted runs. When looking at early precision (P(30)) the variations were large, but slowly disappear for later precision (P(100)). All measures correlate fairly well.

Automatic Retrieval. The category of automatic runs was by far the most common category for result submissions. 79 out of the 100 submitted runs were in this category. In Table 1 the best run of each participating system per category is shown.

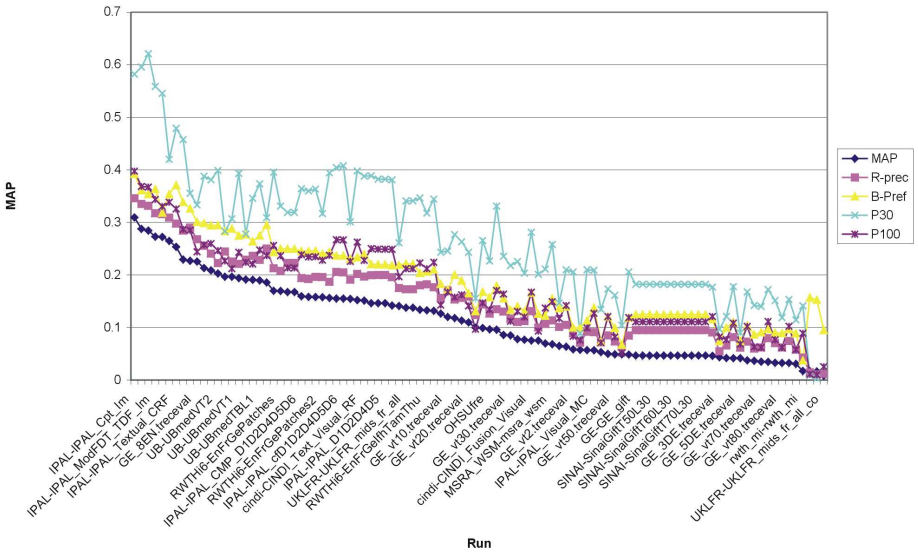


Fig. 3. Results for best runs of each system in each category, ordered by MAP

We can see that the best submitted automatic run was a mixed run and that other mixed runs had very good results. Nonetheless, several of the very good results were textual only, so a generalisation does not seem completely possible. Visual systems had a fairly low overall performance, although for the ten visual topics, their performance was very good.

Manual and Interactive Retrieval. Figure 2 shows the submitted manual and interactive runs. With the small numbers of manual runs, generalisation is

Table 1. Overview of the automatic runs

Run identifier	visual	textual	MAP	R–Prec
IPAL_Cpt_Im	x	x	0.3095	0.3459
IPAL_Textual_CDW		x	0.2646	0.3093
GE_8EN.treceval		x	0.2255	0.2678
UB-UBmedVT2	x	x	0.2027	0.2225
UB-UBmedT1		x	0.1965	0.2256
UKLFR_origmids_en_en		x	0.1698	0.2127
RWTHi6-EnFrGePatches	x	x	0.1696	0.2078
RWTHi6-En		x	0.1543	0.1911
OHSU_baseline_trans		x	0.1264	0.1563
GE_vt10.treceval	x	x	0.12	0.1703
SINAI-SinaiOnlytL30		x	0.1178	0.1534
CINDI_Fusion_Visual	x		0.0753	0.1311
MSRA_WSM-msra_wsm	x		0.0681	0.1136
IPAL_Visual_SPC+MC	x		0.0634	0.1048
RWTHi6-SimpleUni	x		0.0499	0.0849
SINAI-SinaiGiftT50L20	x	x	0.0467	0.095
GE-GE_gift	x		0.0467	0.095
UKLFR_mids_en_all_co	x	x	0.0167	0.0145

Table 2. Overview of the manual and interactive runs

Run identifier **manual**	visual	textual	MAP	R-Prec
OHSUeng		x	0.2132	0.2554
IPAL_CMP_D1D2D4D5D6	x		0.1596	0.1939
INSA-CISMef	x	x	0.0531	0.0719
Run identifier interactive	**visual**	**textual**	**MAP**	**R–Prec**
IPAL_Textual_CRF		x	0.2534	0.2976
OHSU-OHSU_m1	x	x	0.1563	0.187
CINDI_Text_Visual_RF	x	x	0.1513	0.1969
CINDI_Visual_RF	x		0.0957	0.1347

difficult. The first interactive run had good performance but was still not better than the best automatic run of the same group.

2.7 Conclusions

The best overall run by the IPAL institute is an automatic run using visual and textual features. We can tell from the submitted runs that interactive and manual runs do not perform better than the automatic runs. This may be partly due to the fact that most groups submitted more automatic runs than other runs. The automatic approach appears to be less time–consuming and most research groups have more experience in optimising these runs. Visual features seem to be mainly good for the visual topics. Text–only runs perform very well, and only a few mixed runs manage to be better.

3 The Medical Automatic Annotation Task

Automatic image annotation is a classification task, where a given image is automatically labelled with a text describing its contents. In restricted domains, the annotation may be just a class from a constrained set of classes or it may be an arbitrary narrative text describing the contents of the images. In 2005, the medical automatic annotation task was performed in ImageCLEF to compare state–of–the–art approaches to automatic image annotation and classification [11]. This year's medical automatic annotation task builds on top of last year: 1,000 new images were collected and the number of classes more than doubled, resulting in a harder task.

3.1 Database and Task Description

The complete database consists of 11,000 fully classified medical radiographs taken randomly from medical routine at the RWTH Aachen University Hospital. 9,000 of these were release together with their classification as training data, another 1,000 were also published with their classification as validation data to allow for tuning classifiers in a standardised manner. One thousand additional images were released at a later date without classification as test data. These 1,000 images had to be classified using the 10,000 images (9,000 training + 1,000 validation) as training data. The complete database of 11,000 images was

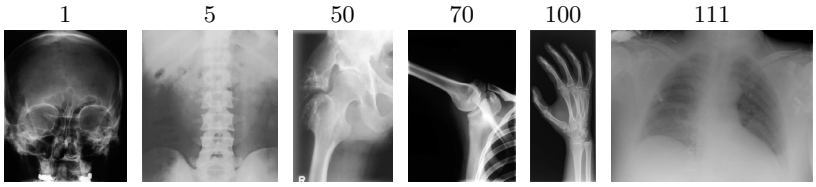


Fig. 4. Example images from the IRMA database with their class numbers

subdivided into 116 classes according to the IRMA code [12]. The IRMA code is a multi-axial code for the annotation of medical images. Currently, this code is available in English and German.

Example images from the database together with class numbers are given in Figure 4. The classes in the database are not uniformly distributed: class 111 has a 19.3% share of the complete dataset, class 108 has a 9.2% share, and 6 classes have only 1‰ or less.

3.2 Participating Groups and Methods

27 groups registered and 12 of these submitted runs. Here, a short description of the methods of the submitted runs is provided. Groups are listed alphabetically by their group ID, which is later used in the results section to refer to the groups.

- *CINDI.* The CINDI group from Concordia University in Canada submitted 5 runs using a variety of features including MPEG–7 Edge Histogram Descriptor, MPEG–7 Color Layout Descriptor, invariant shape moments, downscaled images, and semi–global features. Some experiments combine these features with a PCA transformation. For 4 of the runs, a support vector machine (SVM) is used for classification with different multi–class voting schemes; in one run, the nearest neighbour decision rule is applied.
- *DEU.* The group of the Dokuz Eylul University in Tinaztepe, Turkey submitted one run which uses the MPEG–7 Edge Histogram as image descriptor and a 3–nearest neighbour classifier for classification.
- *MedIC-CISMeF.* The team from the INSA Rouen, France submitted 4 runs. Two use a combination of global and local image descriptors and the other two use local image descriptors. Features are dimensionality reduced by PCA and runs using the same features differing in the PCA coefficients are kept. The local features include statistical measures extracted from image regions and texture information. Yielding a 1953–dimensional feature vector when only local features are used and 2074–dimensional when local and global features are combined. For classification a SVM with RBF kernel is used.
- *MSRA.* The Web Search and Mining Group from Microsoft Research Asia submitted two runs. One run uses a combination of gray-block features, block-wavelet features, features accounting for binarised images, and an edge histogram. In total a 397-dimensional feature vector is used. The other run

uses a bag of features approach with vector quantisation where a histogram of quantised vectors is computed region-wise on the images. In both runs, SVMs are used for classification.

- *MU I2R.* The Media Understanding group of the Institute for Infocomm Research, Singapore submitted one run. A two–stage medical image annotation method was applied. First, the images are reduced to 32x32 pixels and classified using a SVM. Then, the decisions where the SVM was not sure, the decision was refined using a classifier that was trained on a subset of the training images. In addition to downscaled images, SIFT features and PCA transformed features were used for classification.
- *NCTU DBLAB.* The DBLAB of the National Chiao Tung University in Hsinchu, Taiwan submitted one run using tree image features, Gabor texture features, coherence moment and related vector layout as image descriptors. The classification was done using a nearest neighbour classifier.
- *OHSU.* The Department of Medical Informatics & Clinical Epidemiology of the Oregon Health and Science University in Portland, OR, USA submitted 4 runs. For image representation, a variety of descriptors was tested including 16×16 pixel versions of the images, and partly localised GLCM features. For classification multilayer perceptrons were used and settings were optimised using the development set.
- *RWTHi6.* The Human Language Technology and Pattern Recognition Group from the RWTH Aachen, Germany submitted 3 runs. One uses the image distortion model that was used for the best run of last year, and the other a sparse histogram of image patches and absolute position. The image distortion model run uses a nearest neighbour classifier, one of the other runs uses SVMs, and the other uses a maximum entropy classifier.
- *RWTHmi.* The IRMA group of the Medical Informatics Division of the RWTH Aachen University Hospital, Germany submitted two runs which using cross–correlation on 32x32 images with explicit translation shifts, image deformation model for Xx32 images, global texture features as proposed by Tamura, and global texture features as proposed by Castillo based on fractal concepts. For classification a nearest neighbour classifier was used. Weights for these features were optimised on the development set. One of these runs reflects their exact setup from 2005 for comparison.
- *UFR.* The Pattern Recognition and Image Processing group from the University Freiburg, Germany submitted two runs using gradient–like features extracted over interest points. Gradients over multiple directions and scale are calculated and used as a local feature vector. Features are clustered to form a codebook of size 20 and a cluster–cooccurence matrix is computed over multiple distance ranges and multiple angle ranges (since rotation invariance is not desired), resulting in a 4D array per image which is flattened and used as the final feature vector. Classification is done using multi–class SVM in a one–vs–rest approach with a histogram intersection kernel.
- *ULG.* The group from the University of Liège, Belgium extracts a large number of possibly overlapping, square sub–windows of random sizes and at random positions from training images. Then, an ensemble model composed of

Table 3. Results of medical automatic annotation task; expected best marked with '*'

rank	Group	Runtag	Error rate [%]
* 1	RWTHi6	SHME	16.2
* 2	UFR	UFR-ns-1000-20x20x10	16.7
3	RWTHi6	SHSVM	16.7
4	MedIC-CISMeF	local+global-PCA335	17.2
5	MedIC-CISMeF	local-PCA333	17.2
6	MSRA	WSM-msra-wsm-gray	17.6
* 7	MedIC-CISMeF	local+global-PCA450	17.9
8	UFR	UFR-ns-800-20x20x10	17.9
9	MSRA	WSM-msra-wsm-patch	18.2
10	MedIC-CISMeF	local-PCA150	20.2
11	RWTHi6	IDM	20.4
* 12	RWTHmi	rwth-mi	21.5
13	RWTHmi	rwth-mi-2005	21.7
* 14	CINDI	cindi-svm-sum	24.1
15	CINDI	cindi-svm-product	24.8
16	CINDI	cindi-svm-ehd	25.5
17	CINDI	cindi-fusion-KNN9	25.6
18	CINDI	cindi-svm-max	26.1
* 19	OHSU	OHSU-iconGLCM2-tr	26.3
10	OHSU	OHSU-iconGLCM2-tr-de	26.4
21	NCTU	dblab-nctu-dblab2	26.7
22	MU	I2R-refine-SVM	28.0
23	OHSU	OHSU-iconHistGLCM2-t	28.1
* 24	ULG	SYSMOD-RANDOM-SUBWINDOWS-EX	29.0
25	DEU	DEU-3NN-EDGE	29.5
26	OHSU	OHSU-iconHist-tr-dev	30.8
27	UTD	UTD	31.7
28	ULG	SYSMOD-RANDOM-SUBWINDOWS-24	34.1

20 extremely randomised trees is automatically built based on size–normalised versions of the sub–windows, and operating directly on the pixel values to predict classes of sub–windows. Given this classifier a new image is classified by classifying sub–windows and combining classification decisions.

– *UTD.* The University of Texas, Dallas, USA submitted one run. Images are scaled to 16×16 pixels and their dimensionality is reduced by PCA. A weighted k–nearest neighbour algorithm is applied for classification.

3.3 Results

The results from the evaluation are given in Table 3. The error rates range from 16.2% to 34.1%. Based on the training data, a system guessing the most frequent group for all 1,000 test images would result with 80.5% error rate, since 195 radiographs of the test set were from class 111, which is the biggest class in the training data. A more realistic baseline is given by a nearest neighbour classifier using Euclidean distance to compare the images scaled to 32×32 pixels [13]. This classifier yields an error rate of 32.1%.

The average confusion matrix over all runs is given in Figure 5. It can clearly be seen that a diagonal structure is reached and thus that on the average most of the images were classified correctly. Some classes have high inter–class similarity: in particular classes 108 to 111 are often confused and in total many images from other classes were classified to be from class 111. Obviously, not all classes are

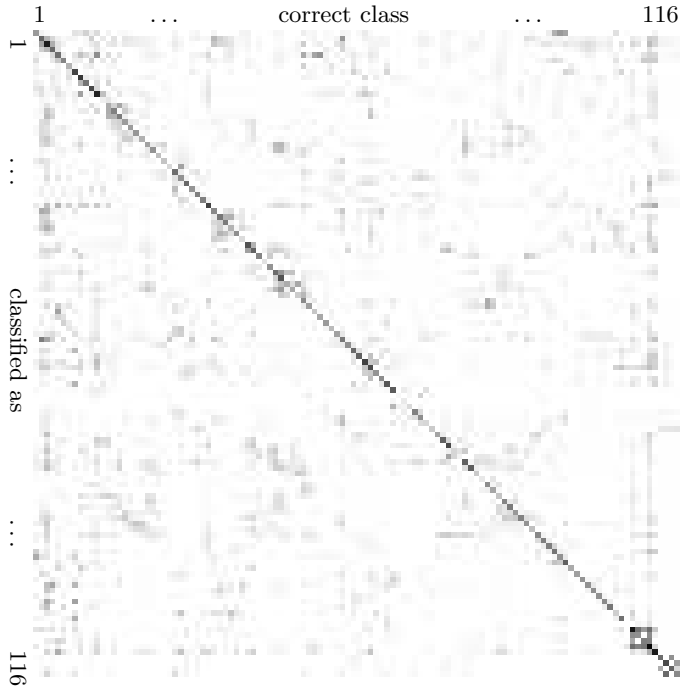


Fig. 5. Average confusion matrix over all runs. Dark points denote high entries, white points zero. X–axis shows the correct class and the Y-axis the class to which images were classified. Values are in logarithmic scale.

equally difficult, a tendency that classes with only few training instances are harder to classify than classes with a large amount of training data can be seen.

3.4 Discussion

The most interesting observation can be seen when comparing the results with those of last year: The RWTHi6–IDM [14] system that performed best in last years task (error rate: 12.1%) obtained an error rate of 20.4%. This increase in error rate can be explained by the larger number of classes and thus more similar classes that can easily be confused, on the other hand, 10 methods clearly outperform this result, 9 of these use SVMs as classifier (ranks 2-10) and one uses a discriminatively trained log–linear model (rank 1). Thus, it can clearly be said, that the performance of image annotation techniques strongly improved over one year, and that techniques that were initially developed for object recognition and detection are very well suited for the automatic annotation.

Given the confidence files of all runs, we tried to combine the classifiers by the sum rule. Therefore, all confidence files were normalised such that the confidences could be interpreted as a–posteriori probabilities $p(c|x)$ where c is the class and x

the observation. Unlike last year, where this technique could not improve results, clear improvements are possible combining several classifiers [15]: Using the top 3 ranked classifiers in combination, an error rate of 14.4% was obtained. The best result is obtained combining the top 7 ranked classifiers. No parameters were tuned but classifiers were combined equally.

4 Overall Conclusions

For the retrieval task, none of the manual or interactive techniques were significantly better than automatic runs. The best–performing systems used visual and textual techniques combined, but several times a combination of visual and textual features did not improve a system's performance. Thus, combinations for multi-modal retrieval need to done carefully. Purely visual systems only performed well on the visual topics. For the automatic annotation task, discriminative methods outperformed methods based on nearest neighbour classification and the top–performing methods were based on the assumption that images consist of image parts, which can be modelled more or less independently.

One goal for future tasks is to motivate groups to work more on interactive and manual runs. Given enough manpower, such runs should be better than optimised automatic runs. Another future goal is to motivate an increasing number of subscribed groups to participate. Collections are planned to become larger. As some groups already complained about too large datasets for their computationally very expensive methods, a smaller database might be provided as an option for these groups to at least submit some results and compare them with other techniques. For the automatic annotation task, one goal is to use text labels with varying annotation precision rather than simple class–based annotation and to consider semi–automatic annotation methods.

Acknowledgements

This work was partially funded by the DFG (Deutsche Forschungsgemeinschaft) under contracts NE-572/6 and Le-1108/4, the Swiss National Science Foundation (FNS) under contract 205321-109304/1, the American National Science Foundation (NSF) with grant ITR–0325160, and the EU 6th Framework Program with the SemanticMining (IST NoE 507505) and MUSCLE NoE.

References

1. Clough, P., Müller, H., Sanderson, M.: Overview of the CLEF cross-language image retrieval track (ImageCLEF) 2004. In: Peters, C., Clough, P.D., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B. (eds.) CLEF 2004. LNCS, vol. 3491, pp. 597–613. Springer, Heidelberg (2005)
2. Clough, P., Grubinger, M., Deselaers, T., Hanbury, A., Müller, H.: Overview of the ImageCLEF 2006 photographic retrieval and object annotation tasks. In: Proceedings of the CLEF 2006 workshop, Alicante, Spain. LNCS (September 2007)

3. Hersh, W., Müller, H., Jensen, J., Yang, J., Gorman, P., Ruch, P.: ImageCLEFmed: A text collection to advance biomedical image retrieval. Journal of the American Medical Informatics Association, 488–496 (2006)
4. Deselaers, T., Müller, H., Clogh, P., Ney, H., Lehmann, T.M: The CLEF 2005 automatic medical image annotation task. CLEF 2005 (2006)
5. Rosset, A., Müller, H., Martins, M., Dfouni, N., Vallée, J.-P., Ratib, O.: Casimage project – a digital teaching files authoring environment. Journal of Thoracic Imaging 19, 1–6 (2004)
6. Candler, C.S., Uijtdehaage, S.H., Dennis, S.E.: Introducing HEAL: The health education assets library. Academic Medicine 78, 249–253 (2003)
7. Wallis, J.W., Miller, M.M., Miller, T.R., Vreeland, T.H.: An internet–based nuclear medicine teaching file. Journal of Nuclear Medicine 36, 1520–1527 (1995)
8. Glatz–Krieger, K., Glatz, D., Gysel, M., Dittler, M., Mihatsch, M.J.: Webbasierte Lernwerkzeuge für die Pathologie – web–based learning tools for pathology. Pathologe 24, 394–399 (2003)
9. Hersh, W., Jensen, J., Müller, H., Gorman, P., Ruch, P.: A qualitative task analysis of biomedical image use and retrieval. In: ImageCLEF/MUSCLE workshop on image retrieval evaluation, Vienna, Austria, pp. 11–16 (2005)
10. Müller, H., Despont–Gros, C., Hersh, W., Jensen, J., Lovis, C., Geissbuhler, A.: Health care professionals' image use and search behaviour. In: Proceedings of Medical Informatics Europe (MIE 2006), Maastricht, Netherlands, pp. 24–32 (2006)
11. Müller, H., Geissbuhler, A., Marty, J., Lovis, C., Ruch, P.: The Use of medGIFT and easyIR for Image CLEF 2005. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, pp. 724–732. Springer, Heidelberg (2006)
12. Lehmann, T.M., Schubert, H., Keysers, D., Kohnen, M., Wein, B.B: The IRMA code for unique classification of medical images. In: Proceedings SPIE, vol. 5033, pp. 440–451 (2003)
13. Keysers, D., Gollan, C., Ney, H.: Classification of medical images using non-linear distortion models. In: Proceedings BVM 2004, Bildverarbeitung für die Medizin, Berlin, Germany, pp. 366–370 (2004)
14. Deselaers, T., Weyand, T., Keysers, D., Macherey, W., Ney, H.: FIRE in ImageCLEF 2005: Combining content-based image retrieval with textual information retrieval. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, pp. 688–698. Springer, Heidelberg (2006)
15. Kittler, J.: On combining classifiers. IEEE Transactions on Pattern Analysis and Machine Intelligence 20, 226–239 (1998)