Overview of the ImageCLEFmed 2006 medical retrieval and annotation tasks

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

[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 medial image retrieval and the medical annotation tasks of ImageCLEF 2006. These tasks are described in a separate paper from the other task to reduce the size of the overview papaer.These two medical tasks are described separately with respect to the goals, databases used, topics created and distributed among participants, results and techniques used. The best performing techniques are described in more detail to provide better insights about successful strategies. Some ideas for future tasks are also presented.

The ImageCLEFmed medical image retrieval task had 12 participating groups and received 100 submitted 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 runs were mixed, i.e., using visual and textual information. None of the manual or interactive techniques were significantly better than those used for the automatic runs. The best–performing systems used visual and textual techniques combined, but combinations of visual and textual features often did not improve a system's performance. Purely visual systems only performed well on the visual topics.

The medical automatic annotation used a larger database in 2006, with 10'000 training images and 116 classes, up from 57 in 2005. Twelve participating groups submitted 27 runs. Despite the much larger number of classes, results were almost as good as in 2005 and a clear improvement in performance could be shown. The best–performing system of 2005 would have only received a position in the upper middle part in 2006.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.3.7 Digital Libraries; H.2.3 [**Database Management**]: Languages—*Query Languages*

General Terms

Measurement, Performance, Experimentation

Keywords

Image Retrieval, Performance Evaluation, Image Classification, Medical Imaging

1 Introduction

ImageCLEF[1] [3] started within CLEF[2] (Cross Language Evaluation Forum) in 2003. A medical image retrieval task was added in 2004 to explore domain–specific multilingual information 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 are 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 found on the task web pages for ImageCLEFmed[3] and the medical annotation task[4]. A detailed analysis of the 2005 medical image retrieval task is available in [8].

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 in a row with the same dataset of over 50'000 images from four distinct 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 categories of being amenable to visual, textual, or mixed retrieval methods.

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 judgments was similar to 2005 and, for the evaluation of the results, the trec_eval package was used, since it is the standard in information retrieval.

2.2 Registration and participation

In 2006, a record number of 47 groups registered for ImageCLEF and among these, 37 also registered for the medical image retrieval task. Groups came from four continents and from a total of 16 countries.

Unfortunately, many of the registered group did not send in results In the end, 12 groups from 8 countries submitted results. Each entry below describes briefly the techniques used for their submissions.

- *Concordia University, Canada.* The CINDI group from Concordia University, Montreal, Canada submitted a total of four runs, one purely textual, one purely visual, and two

[1] http://ir.shef.ac.uk/imageclef/
[2] http://www.clef-campaign.org/
[3] http://ir.ohsu.edu/images
[4] http://www-i6.informatik.rwth-aachen.de/~deselaers/imageclef06/medicalaat.html

combined runs. Text retrieval was based on Apache Lucene. For visual information a combination of global and local features were used and compared using the Euclidean distance. Most of the submissions used relevance feedback.

- *Microsoft Research, China.* Microsoft Research China submitted one purely visual run using a combination of various features accounting for color, texture, and blocks.

- *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, the three languages were separately mapped to UMLS terms and then applied to retrieval. Several classifiers based on SVMs and other classical approaches were used and combined.

- *University Hospitals of Freiburg, Germany.* The Freiburg group submitted a total of 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 query language and combination with GIFT settings.

- *Jaen University (SINAI), Spain.* The SINAI group submitted 12 runs, three of them using only textual information and nine using a textual retrieval system and addind provided data from the GIFT image retrieval system. The runs differed in settings for "information gain" and the weighting of textual and visual information.

- *Oregon Health and Science University (OHSU), USA.* OHSU performed manual modification of queries and then attempted to augment output by fusing results from visual runs. One set of runs from OHSU established a baseline using the text of the topics as given. Another set of runs then manually modified the topic text removing common words and adding synonyms. For both sets of runs, there were submissions in each of the three individual languages (English, French, German) plus a merged run with all three and another run with the English topics expanded with automatic translation using the Babelfish translator. The manual modification of the queries improved performance substantially, though still below other groups' automated methods. 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 high levels of retrieval output, e.g., precision at 10 and 30 images.

- *I2R Medical Analysis Lab, Singapore.* Their submission was together with the IPAL group from the same lab.

- *MedGIFT, University and Hospitals of Geneva, Switzerland.* The University and Hospitals of Geneva relied on two retrieval systems for their submission. The visual part was performed with the medGIFT retrieval system. The textual retrieval used a mapping of the query and document text towards concepts in the MeSH (Medical Subject Headings) terminology. Then, matching was performed with a frequency–based weighting methods 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.* RWTHi6 submitted a total of nine runs, all using the FIRE retrieval system and a variety of features describing color, texture, and global appearance in different ways. For one of the runs, 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, three runs were purely visual, and the remaining five runs used textual and visual information. All runs were fully–automatic runs without any user interaction or manual tuning.

- *RWTH Aachen University – Medical Informatics, Germany.* RWTHmi submitted two purely visual runs without any user interaction. Both runs used a combination of various global appearance features compared using invariant distance measures and texture features. The runs differed in the weights for the features used.

- *State University New York, Buffalo, USA.* SUNY submitted fourruns, two purely textual and two using textual and visual information. Parameters for their system were tuned using the ImageCLEF 2005 topics, and automatic relevance feedback was used in different variations.

- *LITIS Lab, INSA Rouen, France.* The INSA group from Rouen submitted one run using visual and textual information. For the textual information the MeSH dictionaries were used and the images were represented by various features accounting for global and local information. Most of the topics were treated fully automatic, and only four topics were treated with manual interaction.

2.3 Databases

In 2006, the same dataset was used as in 2005 containing four distinct sets of images. The Casimage[5] dataset was made available to participants [14], containing almost 9'000 images of 2'000 cases [15]. Images present in Casimage include mostly radiology modalities, 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 the PEIR[6] (Pathology Education Instructional Resource) database with annotation based on the HEAL[7] project (Health Education Assets Library, mainly Pathology images [1]). This dataset contains over 33.000 images with English annotations, with the annotation being on a per image and not a per case basis as in Casimage. The nuclear medicine database of MIR, the Mallinkrodt Institute of Radiology[8] [16], was also made available to us for ImageCLEFmed. This dataset contains over 2.000 images mainly from nuclear medicine with annotations provided per case and in English. Finally, the PathoPic[9] collection (Pathology images [5]) was included in our dataset. It contains 9.000 images with extensive annotation on a per image basis 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.

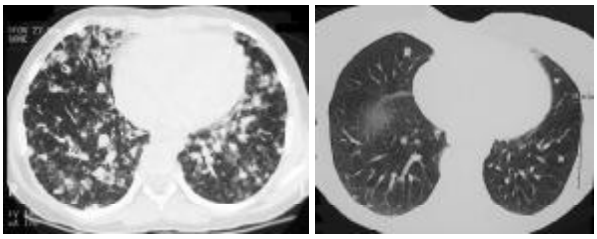
2.4 Query topics

The query topics were based on two surveys performed in Portland and Geneva [7, 12]. In addition to this, a log file of a media search engine HON[10] was used to create topics. Based on the surveys, topics for ImageCLEFmed were developed 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).

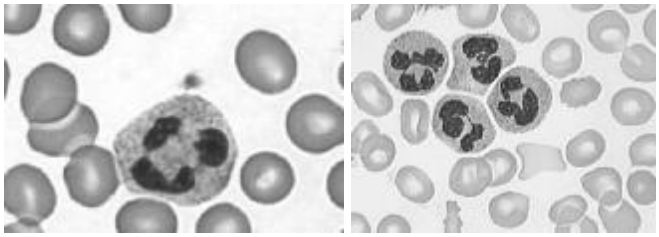
Still, as the HON log-files indicated rather general topics than the fairly specific ones used in 2005, we used real queries from these log-files in 2006. We could not use the most frequent queries, since they were too general, e.g. heart, lung, etc., but rather those that satisfied at least two of the

[5] http://www.casimage.com/
[6] http://peir.path.uab.edu/
[7] http://www.healcentral.com/
[8] http://gamma.wustl.edu/home.html
[9] http://alf3.urz.unibas.ch/pathopic/intro.htm
[10] 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.

Figure 1: Example for a visual topic.



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

Figure 2: Example for a mixed topic.

defined axes and that appeared frequently. After identifying over 50 of such candidate topics, we grouped them into three classes based upon as estimation of what retrieval techniques to which 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 corresponding to the database that contains many pathology images.

After choosing ten queries for each of the three categories, we searched query images on the web manually. In 2005, images were taken partly from the collection. Although they were cropped most of the time, having images from another collection made the visual task more challenging, as these images could be from other modalities and have completely different characteristics concerning texture, luminosity, etc. This year we created 10 topics for each of the 3 groups for a total of 30 topics. Figures 1, 2, 3 show examples for a visual, a mixed and a semantic topic.

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

Figure 3: Example for a semantic topic.

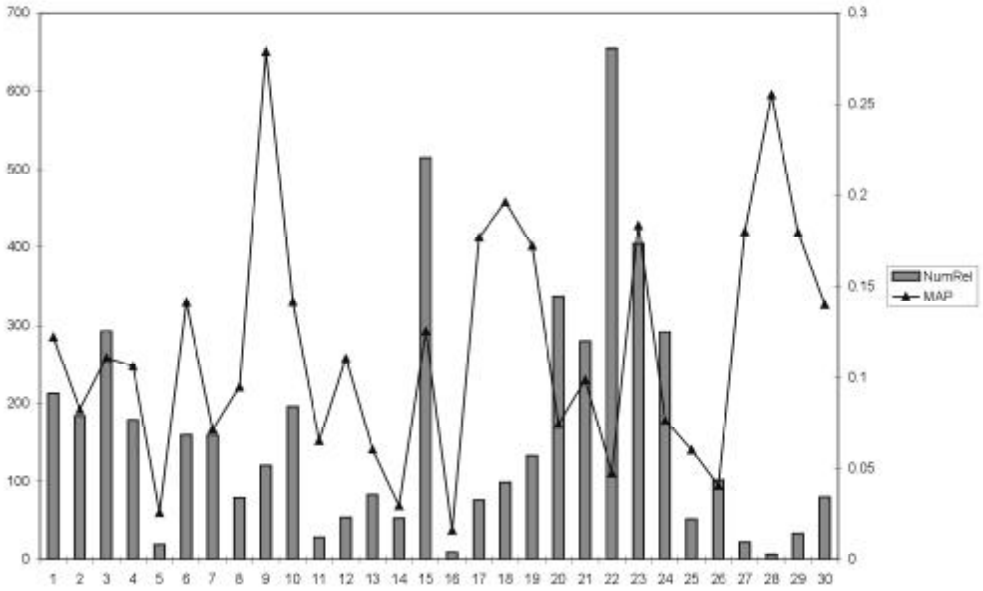


Figure 4: Evaluation results and number of relevant images per topic.

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 of anywhere from 647 to 1187 images, with a mean of 910 per topic. Relevance judgments 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 judgments; ten topics had duplicate judgments that we will analyze later.) The judgments 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.

2.6 Submissions and Results

A total of 12 groups participated in ImageCLEFmed 2006 from eight different countries (Canada, China, France, Germany, Singapore, Spain, Switzerland, and the United States). These groups collectively submitted 100 runs, with each group submitting anywhere 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 the submitted runs were automatic. There were fewer visual runs than there were textual and mixed runs.

Figure 4 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 other only had very few. Likewise, performance could be extremely good for a few topics and extremely bad for others. There does not appear to be a direct connection between number of relevant images

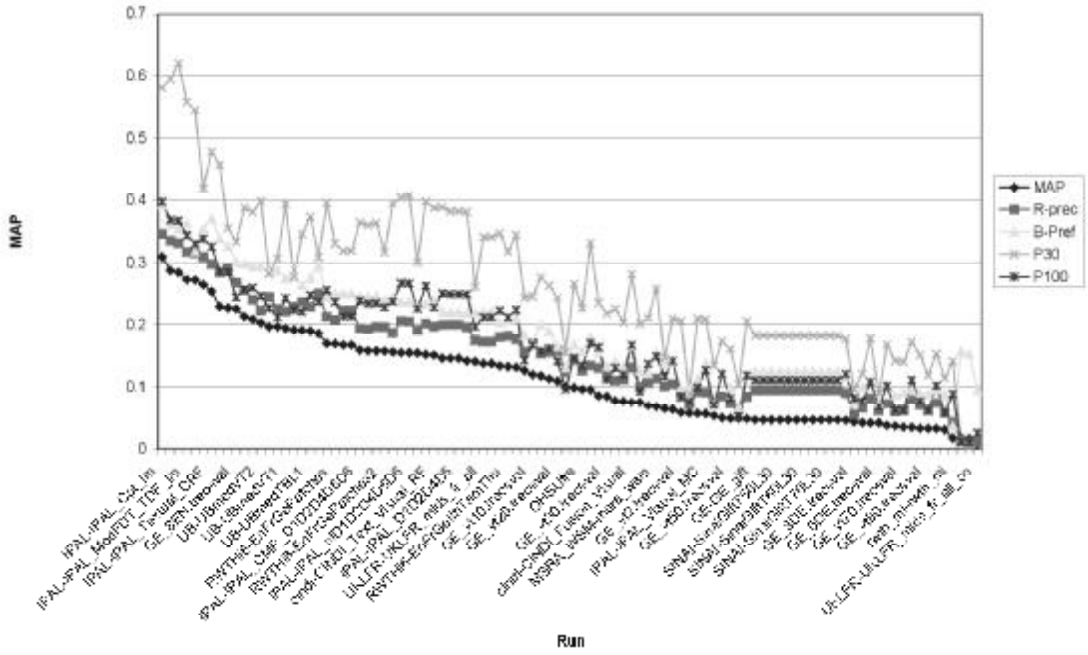


Figure 5: Evaluation results for the best runs of each system in each category, ordered by MAP.

for a topics and the average performance that systems obtain.

In Figure 5 is a comparison of several performance measurements for all submitted runs. In particular when looking at early precision (P(30)) these variations were quite large, but slowly disappear for later precision (P(100)). On the other hand, these measures do seem to correlate fairly well.

2.6.1 Automatic retrieval

The category of automatic runs was by far the most common category for results submissions. A total of 79 of the 100 submitted runs were in this category. In Table 1 the best run of each participating system per category is shown as is in the following tables. Showing all 100 runs would have results in information difficult to read.

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 generalization does not seem completely possible. Visual systems had a fairly low overall performance, although for the first ten visual topics, their performance was very good.

2.6.2 Manual retrieval

Figure 2 shows the submitted manual runs. With the small numbers of these runs, generalization is difficult.

2.6.3 Interactive retrieval

Table 3 shows the submitted interactive runs. The first run had good performance but was still not better than the best automatic run of the same group.

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 runs.

Run identifier	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

Table 3: Overview of the interactive runs.

Run identifier	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

2.7 Conclusions

The best overall run by the IPAL institute is an automatic run using visual and textual features. From the submitted runs, we can say 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 many more automatic runs than other runs. The automatic approach appears to be less time–consuming and most research groups have more experience in optimizing these runs. Visual features seem to be mainly good for the visual topics but fail to help for the semantic features. 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 labeled 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. Last year, the medical automatic annotation task was performed in ImageCLEF to compare state-of-the-art approaches to automatic image annotation and classification and to make a first step toward using automatically annotated images in a multi-modal retrieval system [13]. This year's medical automatic annotation task builds on top of last year, with 1,000 new images to be classified were collected and the number of classes is more than doubled, resulting in a harder task.

3.1 Database & Task Description

The complete database consists of 11,000 fully classified radiographs taken randomly from medical practice at the RWTH Aachen University Hospital. A total of 9,000 of these were released together with their classification as training data, with another 1,000 also published with their classification as validation data to allow the groups for tuning their classifiers in a standardized manner. One thousand additional images were released at a later date without their 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 subdivided into 117 classes according to the complete IRMA code annotation [11]. The IRMA code is a multi-axial scheme assessing anatomy, biosystem, creation and direction of imaging. Currently, this code is available in English and German, but could easily be translated to other languages. It is planned to use the result of such automatic annotation experiments for further, textual image retrieval tasks in the future.

Example images from the database together with their class numbers are shown in Figure 6. The classes in the database are not uniformly distributed, for example, class 111 has a 19.3% share of the complete dataset, class 108 has a 9.2% share of the database, while six classes have only 1‰ or less.

3.2 Participating Groups & Methods

In total, 28 groups registered and 12 of these submitted runs. For each group, a brief description of the methods of the submitted runs is provided. The 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 Montreal, 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 of the experiments combine these features with a principal component analysis (PCA). The dimensionality of the feature vectors is up to 580. For four of the runs, a support vector machine (SVM) is used for classification with different multi-class voting schemes. In one run, the nearest neighbor decision rule is applied. The group expects the run `cindi-svm-sum` to be their best submission.

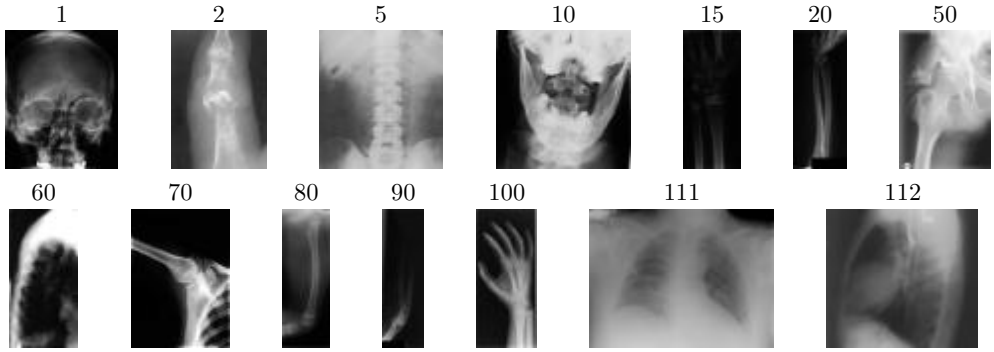


Figure 6: Example images from the IRMA database together with their class numbers. the bottom row emphasized the intra-class variety of the IRMA database.

DEU. The Department of Computer Engineering of the Dokuz Eylul University in Tinaztepe, Turkey submitted one run using the MPEG-7 Edge Histogram as 80-dimensional image descriptor and a 3-nearest neighbor classifier for classification.

MedIC-CISMeF. The CISMeF team from the INSA Rouen in Saint-Étienne-du-Rouvray Cedex, France submitted four runs. Two of them use a combination of global and local image descriptors and the other two are based on local image descriptors only. Features are dimensionality reduced by a PCA and those runs which use the same features differ in the PCA coefficients kept. The features include statistical measures extracted from image regions and texture information. This yields a 1953-dimensional feature vector when only local features are used and 2074 dimensional feature vector when local and global features are combined. These feature vectors are reduced by PCA to 335 and 470 dimensions, respectively. For classification a support vector machine with radial basis function kernel is used. The group expects the run `local+globalPCA450` to be their best submission.

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 binarized 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 quantization, where a histogram of quantized vectors is computed region-wise on the images. In both runs, SVM is used for classification. The group did not identify which of these they expect to be better.

MU I2R. The Media Understanding Group of the Institute for Infocomm Research, Singapore submitted one run. In this run, a two-stage medical image annotation method was applied. In the first stage, the images are reduced to 32×32 pixels (1024 dimensional vector) and classified using a support vector machine. In the second stage, those decisions for which the support vector machine was unsure were refined using a classifier that was trained on a subset of the training images. In addition to down-scaled images, SIFT (scale invariant feature transformation) features and principal components of 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 neighbor 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 localized gray level cooccurrence matrix (GLCM) features resulting in a feature vector of up to 380 components. For classification, a multilayer perceptron were used and settings were optimized using the development set.

RWTHi6. The Human Language Technology and Pattern Recognition Group from the RWTH Aachen University in Aachen, Germany submitted three runs. One uses the image distortion model (IDM) that was used for the best run of last year, and the other a sparse histogram of image patches and absolute position. The IDM run is based on a nearest neighbor classifier, while the other runs use SVM or a maximum entropy classifier. The feature vectors for the IDM experiments have less than 1024 components and the sparse histograms have 65536 bins. The group expects the run `SHME` to be their best submission.

RWTHmi. The Image Retrieval in Medical Applications (IRMA) group, Department of Medical Informatics, RWTH Aachen University Hospital in Aachen, Germany submitted two runs using cross-correlation on 32×32 images with explicit translation shifts, IDM for X×32 images, global texture features as proposed by Tamura, and global texture features as proposed by Castelli et al. based on fractal concepts resulting in an approximately 2500-dimensional feature vector. For classification, a nearest neighbor classifier was used. For the run `RWTHmi-opt` weights for these features were optimized on the development set, and for the run `RWTHmi-baseline` the default parameters of the IRMA system were used.

UFR. The Pattern Recognition and Image Processing group from the University of Freiburg in 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. The features are clustered to form a code book of size 20 and a cluster co-occurrence matrix is computed over multiple distance ranges and multiple angle ranges (since rotation invariance is not desired), resulting in a 4-D array per image which is flattened and used as final, approximately 160000-dimensional, feature vector. Classification is done using multi-class SVM in a one-vs-rest approach with a histogram intersection kernel.

ULG. The Systems and Modeling group of the Institute Montefiore from Liège, Belgium extracts a large number of possibly overlapping, squared sub-windows of random sizes and at random positions from training images. Then, an ensemble model composed by twenty randomized trees is automatically built based on size-normalized versions of the sub-windows. It is operated directly on their pixel values to predict classes of sub-windows. Given this sub-window classifier, a new image is classified via sub-windows and combining the classification decisions. The feature vectors are 576-dimensional. The group expects the run `ULG-SYSMOD-RANDOM-SUBWINDOWS-EX` to be their best submission.

UTD. The Data Mining Laboratory group of the University of Texas at Dallas, Richardson, TX, USA submitted one run. The images are scaled to 16×16 pixels, and their dimensionality is reduced by PCA, resulting in a maximally 256-dimensional feature vector. Then, a weighted k-nearest neighbor algorithm is applied for classification.

MedGIFT. The medGIFT group of the University and Hospitals of Geneva submitted three runs to the medical automatic annotation task. One was entirely based on tf/idf weighting of the GNU Image Finding Tool (GIFT) and thus acted as a baseline using only collection frequencies of features with no learning on the training data supplied. For the second run features are weighted with an additional factor, learned from the supplied training data. For these submissions a 5-NN was used as classifier. The third submission is a combination of several separate runs by

voting. The combined results are quite different, so the combination-run is expected to be the best submission. The runs were submitted after the evaluation ended and are thus not ranked.

3.3 Results

The results from the evaluation are shown in Table 4. The error rates ranged 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 in a 80.5% error rate, since 195 radiographs of the test set were from class 111, which was the biggest class in the training data. A more realistic baseline is given by a nearest neighbor classifier using Euclidean distance to compare the images scaled to 32×32 pixels [9]. This classifier yields an error rate of 32.1%. The average confusion matrix of all submitted runs is shown in Figure 7. Obviously, a diagonal structure is reached. Thus on the average, many images were classified correctly, but it can also be seen that some classes have high inter-class similarity: e.g. classes 108 to 111 are often confused. In total, many images from other classes were classified to be from class 111, which was the class with the highest amount of training data. Obviously, not all classes were equally difficult. A tendency that classes with only few training instances were harder to classify than classes with a large amount of training data could be seen; which was to be expected and had been reported in the literature earlier [6].

Given the confidence files of all runs, we tried to combine the classifiers by the sum rule. Therefore, all confidence files were normalized such that the confidences could be interpreted as a-posteriori probabilities $p(c|x)$ where c was the class and x the observation. Unlike last years results, where this technique could not improve the results, clear improvements were possible combining several classifiers [10]: Using the top 3 ranked classifiers in combination, an error rate of 14.4% was obtained. The best result was obtained combining the top 7 ranked classifiers. Note, that here no additional parameters were tuned but the classifiers were combined weighted equally.

3.4 Discussion

The most interesting observation of this year's evaluation can be seen when comparing the results with the results of last year: The RWTHi6-IDM [4] system that performed best in last years task (error rate: 12.1%) obtained an error rate of 20.4% this year. 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 outperformed this result this year, nine of these use SVMs as a classifier (ranks 2-10) and one using a discriminatively trained log-linear model (rank 1). Thus, it can clearly be stated that the performance of image annotation techniques strongly improved over the last year, and that techniques that were initially developed in the field of object recognition and detection are very well suited for the automatic annotation of medical radiographs.

Another interesting observation drawn from the combination of classifiers was that in contrast to last year, where a combination of arbitrary classifiers from the evaluation did not lead to an improvement over the best submission, this year a clear improvement was obtained by combining several submissions. A reason for this might be the improved performance of the submissions or the higher diversity among the submitted methods.

To give an approximate idea of runtime and memory requirements of the various methods we give the dimensionality of the feature vectors used by the groups. Naturally, the dimension of the feature vectors alone does not say very much about runtime, because the used models for classification have a high impact on runtime and memory consumption, too. However, a trend that high dimensional feature vectors lead to good results can clearly be seen in the results as the best three methods use feature vectors of very high dimensionality (65,536 and 160,000 respectively) and implicitly transform these into even higher dimensional feature spaces by the use of kernel methods.

Table 4: Results of medical automatic annotation task. If a group submitted several runs, the run that was expected to be their best is 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	opt	21.5
	13	RWTHmi	baseline	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
	20	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
	-	medGIFT	combination	29.7
	26	OHSU	OHSU-iconHist-tr-dev	30.8
	-	medGIFT	fw-bwpruned	31.7
	27	UTD	UTD	31.7
	-	medGIFT	baseline	32.0
	28	ULG	SYSMOD-RANDOM-SUBWINDOWS-24	34.1

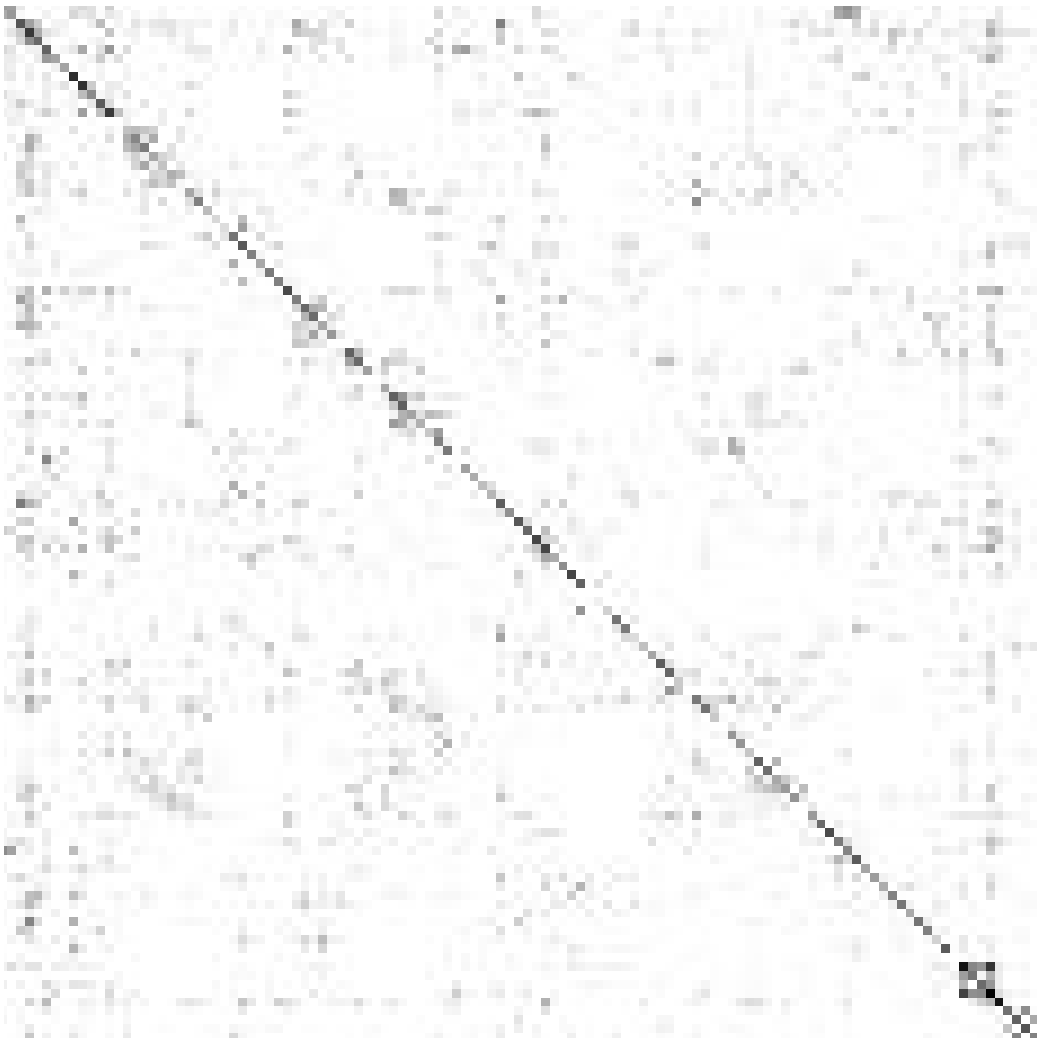


Figure 7: Average confusion matrix over all runs of the medical automatic annotation task. Dark points denote high entries, white points denote zero. On the x-axis, the correct class is given and on the y-axis the class to which images have been classified is given. For visualization purposes values are in logarithmic scale.

4 Overall Conclusions

For the medical retrieval task, none of the manual or interactive techniques were significantly better than those used for the 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 neighbor classification and the top-performing methods were based on the assumption that images consist of images parts which can be modelled more or less independently.

One goal for future tasks is to motivate groups to work more on interactive or manual runs than automated retrieval. With proper manpower, such runs should be better than even optimized automatic runs. Another future goal is to motivate an increasing number of subscribed groups to participate. Collections are planned to become larger as well to stay realistic. Some groups already complained about too large datasets, so a smaller second dataset might be an option for these groups to at least submit some results and compare them with the other techniques.

For the automatic annotation task, a future goal is to use textual labels with varying annotation precision rather than a simple class-based annotation scheme and to consider semi-automatic annotation methods.

Acknowledgements

We would like to thank the CLEF campaign for supporting the ImageCLEF initiative. Furthermore, the authors would like to thank LTUtech[11] for providing the database for the non-medical automatic annotation task and to Tobias Weyand for creating the web interface for submissions.

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 Sixth Framework Program with the SemanticMining project (IST NoE 507505) and the MUSCLE NoE.

References

[1] C. S. Candler, S. H. Uijtdehaage, and S. E. Dennis. Introducing HEAL: The health education assets library. *Academic Medicine*, 78(3):249–253, 2003.

[2] P Clough, M Grubinger, T Deselaers, A Hanbury, and H. Müller. Overview of the ImageCLEF 2006 photo retrieval and object annotation tasks. In *CLEF working notes*, Alicante, Spain, Sep. 2006.

[3] Paul Clough, Henning Müller, and Mark Sanderson. Overview of the CLEF cross–language image retrieval track (ImageCLEF) 2004. In Carol Peters, Paul D. Clough, Gareth J. F. Jones, Julio Gonzalo, M. Kluck, and B. Magnini, editors, *Multilingual Information Access for Text, Speech and Images: Result of the fifth CLEF evaluation campaign*, Lecture Notes in Computer Science, Bath, England, 2005. Springer–Verlag.

[4] Thomas Deselaers, Tobias Weyand, Daniel Keysers, Wolfgang Macherey, and H. Ney. FIRE in ImageCLEF 2005: Combining content-based image retrieval with textual information retrieval. In *Workshop of the Cross–Language Evaluation Forum (CLEF 2005)*, Lecture Notes in Computer Science, page in press, Vienna, Austria, September 2005.

[11] `http://www.ltutech.com/`

[5] K Glatz-Krieger, D. Glatz, M. Gysel, M. Dittler, and M. J. Mihatsch. Webbasierte Lernwerkzeuge für die Pathologie – web–based learning tools for pathology. *Pathologe*, 24:394–399, 2003.

[6] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, Berlin, 2001.

[7] William Hersh, Jeffery Jensen, Henning Müller, Paul Gorman, and Patrick Ruch. A qualitative task analysis of biomedical image use and retrieval. In *ImageCLEF/MUSCLE workshop on image retrieval evaluation*, pages 11–16, Vienna, Austria, September 2005.

[8] William Hersh, Henning Müller, Jeffery Jensen, Jianji Yang, Paul Gorman, and Patrick Ruch. Imageclefmed: A text collection to advance biomedical image retrieval. *Journal of the American Medical Informatics Association*, September/October, 2006.

[9] Daniel Keysers, Christian Gollan, and Hermann Ney. Classification of medical images using non-linear distortion models. In *Proc. BVM 2004, Bildverarbeitung für die Medizin*, pages 366–370, Berlin, Germany, March 2004.

[10] J. Kittler. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, March 1998.

[11] Thomas M. Lehmann, Henning Schubert, Daniel Keysers, M Kohnen, and Berthold B Wein. The irma code for unique classification of medical images. In *Proceedings SPIE*, number 5033, pages 440–451, 2003.

[12] Henning Müller, Christelle Despont-Gros, William Hersh, Jeffery Jensen, Christian Lovis, and Antoine Geissbuhler. Health care professionals' image use and search behaviour. In *Proceedings of the Medical Informatics Europe Conference (MIE 2006)*, Maastricht, The Netherlands, August 2006.

[13] Henning Müller, Antoine Geissbuhler, Johan Marty, Christian Lovis, and Patrick Ruch. The Use of medGIFT and easyIR for ImageCLEF 2005. In *Proceedings of the Cross Language Evaluation Forum 2005*, LNCS, page in press, Vienna, Austria, September 2006.

[14] Henning Müller, Antoine Rosset, Jean-Paul Vallée, Francois Terrier, and Antoine Geissbuhler. A reference data set for the evaluation of medical image retrieval systems. *Computerized Medical Imaging and Graphics*, 28:295–305, 2004.

[15] Antoine Rosset, Henning Müller, Martina Martins, Natalia Dfouni, Jean-Paul Vallée, and Osman Ratib. Casimage project – a digital teaching files authoring environment. *Journal of Thoracic Imaging*, 19(2):1–6, 2004.

[16] J. W. Wallis, M. M. Miller, T. R. Miller, and T. H. Vreeland. An internet–based nuclear medicine teaching file. *Journal of Nuclear Medicine*, 36(8):1520–1527, 1995.