# 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 will describe the medial image annotation and medical image retrieval tasks of ImageCLEF 2006. For the first time these tasks will be described in a separate paper to reduce the size of the ImageCLEF overview, as ImageCLEF in 2006 contains a total of 4 subtasks, each with several participants. The two tasks will be described separately with respect to goals of the task, databases used, topics created and distribute among participants, results and techniques used. The best–performing techniques will be described in a little more detail, to get better ideas about successful strategies. Some ideas for future tasks will also be presented.

The ImageCLEFmed medical image retrieval task had 12 participating groups and received a total of 100 submitted runs. Most runs were automatic runs and only few manual and feedback. Purely textual runs were in majority compared to purely visual runs but most runs were mixed, so using visual and textual information. None of the manual or feedback techniques was significantly better than the automatic runs, which might be linked to the fact that manual/feedback runs require more work with respect to user interaction and many groups prefer optimising automatic runs rather than spending work on user interactions. The best–performing systems used visual and textual information combined but several times a combination of visual and textual features did not improve a system's performance. Purely visual systems only perform well on the ten visual topics.

The medical automatic annotation used a larger database in 2006, with 10'000 training images of 116 classes instead of 57 classes in 2005. 12 participating groups submitted in total 27 runs. Despite the much larger number of classes, results stayed 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 Managment**]: Languages—*Query Languages*

## General Terms

Measurement, Performance, Experimentation

## Keywords

Image Retrieval, Performance Evaluation, Image Classification, Medical Imaging

# 1 Introduction

ImageCLEF[1] [3] has 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 and also multimodal retrieval by combining visual and textual features for retrieval. Since 2005, a medical retrieval and a medical image annotation task are both presented as part of ImageCLEF.

This paper is concentrating on the two medical tasks whereas a second paper [2] will describe the new object classification task and the new 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 [7].

# 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 topics development was slightly enlarged by using the log files of a medical media search engine of the health on the net (hon) foundation. This analysis shows a need for more general topics not covering the entire four axes defined in 2005:

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

Other than that, the process of relevance judgements was similar to 2005 and for the evaluation of the results the trec_eval package was used as 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, several of the submitting groups did only register but finally did not manage to send in results, which is a common phenomenon in benchmarking events, as the data alone even

[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

without participation is valuable for research. A survey is currently being performed to find out more about the reasons for not participating after a registration. First responses indicate a lack of time and computing resources for large databases. Some people said to not have a working system and using the test data for system development. All said that they plan to re–inscribe for ImageCLEFmed to participate in future tasks. In 2005, a similar survey was performed and the result was that many groups did not have the manpower for a proper evaluation.

25 groups from 14 countries registered but finally did not manage to submit runs for the evaluation

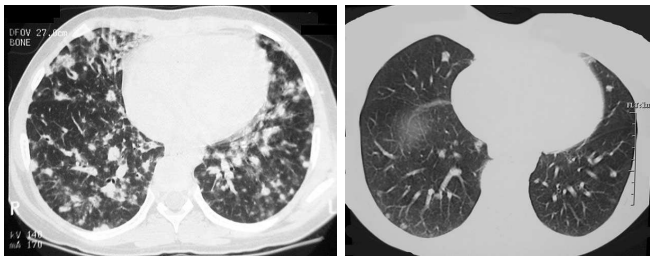
The following groups submitted results. Each entry also describes in a few words the techniques used for their submissions.

- *Concordia University, Canada.*
- *Microsoft Research, China.*
- *Institute for Infocomm Research I2R–IPAL, Singapore.*
- *University Hospitals of Freiburg, Germany.*
- *Jain University (SINAI), Spain.*
- *Oregon Health and Science University (OHSU), USA.*
- *I2R Medical Analysis Lab, Singapore.* Their submission was together with the IPAL group from the the 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 war 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 method. All results were automatic runs using visual, textual and mixed features. Separate runs were submitted for the three languages.
- *RWTH Aachen – Computer Science, Germany.* The visual retrieval was based on the Fire retrieval system using a variety of features.
- *RWTH Aachen – Medical Informatics. Germany.*
- *State University New York, Buffalo, USA.*
- *LITIS Lab, INSA Rouen, France.* The INSA group from Lyon only submitted one visual run

## 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 [13], containing almost 9'000 images of 2'000 cases [14]. 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

[5] http://www.casimage.com/
[6] http://peir.path.uab.edu/
[7] http://www.healcentral.com/



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.

medicine database of MIR, the Mallinkrodt Institute of Radiology[8] [15], 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 into 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

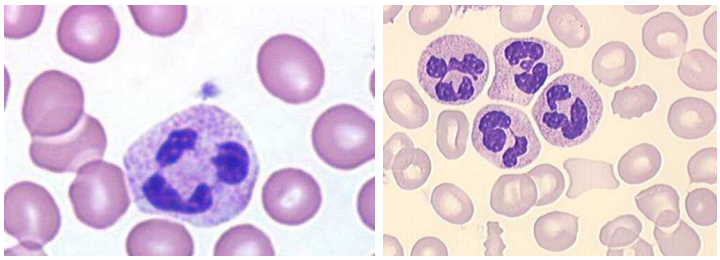
The query topics were based on two surveys performed in Portland and Geneva [6, 11]. In addition to this, a log file of a media search engine of the health on the net (hon[10]) were used to create topics. Based on the surveys, topics for ImageCLEFmed were developed along the following axes:

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

Still, as the hon logfiles indicate rather general topics than the fairly specific ones used in 2005, we used in 2006 real queries from these logfiles. We could not use the most frequent queries (too general: heart, lung, ...) but rather those that satisfy at least two of the defined axes and that appear frequently. After identifying over 50 of such query candidates, we grouped them into several classes (visual, mixed, semantic). 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.

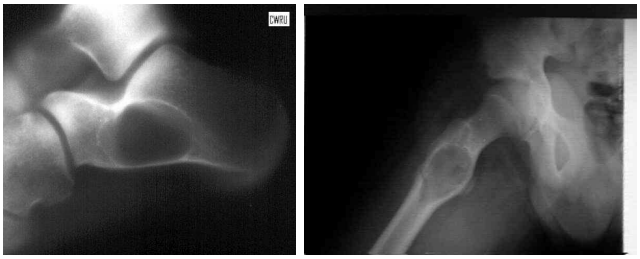
After choosing ten queries for each of the three categories, we were search query images on the web manually. In 2005 images were taken partly from the collection. Although they were cropped most of the time, having images from another collection makes the visual task much hard as these images can be from other modalities and have completely different characteristics concerning 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.

[8] http://gamma.wustl.edu/home.html
[9] http://alf3.urz.unibas.ch/pathopic/intro.htm
[10] http://www.hon.ch/



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.



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.

## 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. This number was keep to under 100 as in 2005 to limit the work of the judges. 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. From these "original" judgements, a qrels file was developed for using trec_eval.

A total of 27,306 relevance judgements were made. (These were primary judgements; ten topics had duplicate judgements that we will analyse later.) There were 62 images that were not judged, a very tiny fraction that we decided to ignore (i.e., as if the image were not in the pool). The judgements were turned into a qrels file, which was then used to calculate results with trec_eval. We took Mean Average Precision (MAP) as a lead measure again, knowing that this is not the most appropriate measure with respect to how a user might judge the retrieval quality.

## 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, interaction – human modification of the query after the output of the system is seen, manual – human modification of the query before the output of the system is seen) and one for the medium used for retrieval (visual, textual and a mixture). The majority of the submitted runs was automatic. Concerning the medium used, there are less visual runs than there

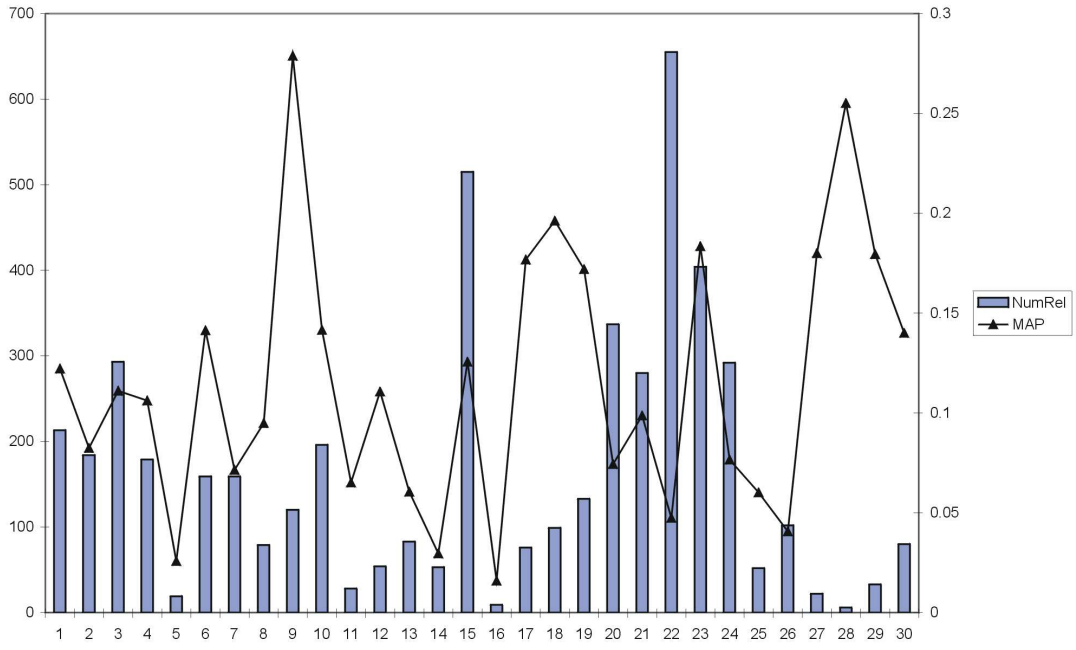



Figure 4: Evaluation results and number of relevant images per topic.

are 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). We can see that the variety in this case is enormous. Some topics have several hundred relevant images in the collection, whereas other only have very few. Likewise, performance can be extremely good for a few topics and extremely bad for others. We can not see a direct connection between number of relevant images for a topics and the average performance that systems obtain.

In Figure 5 we can see a comparison of several performance measures for the system ranking. In particular when looking at early precision (P(30)) these variations can be quite large, but slowly disappear for later precision (P(100)). It can on the other hand be seen fairly well that overall these measure correlate extremely well.

### 2.6.1 Automatic retrieval

The category of automatic runs was by far the most common category for results submissions. 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 is by far a mixed run and other mixed runs have very good results. Still, several of the very good results are textual results, only, so a generalisation does not seem completely possible. Visual systems have a fairly low overall performance but for the first ten visual topics their performance is very good.

## 2.7 Manual retrieval

Figure 2 shows the submitted manual runs. With the small numbers of these runs, no sensible evaluation seems possible, although the textual runs has a better performance than the two other runs.

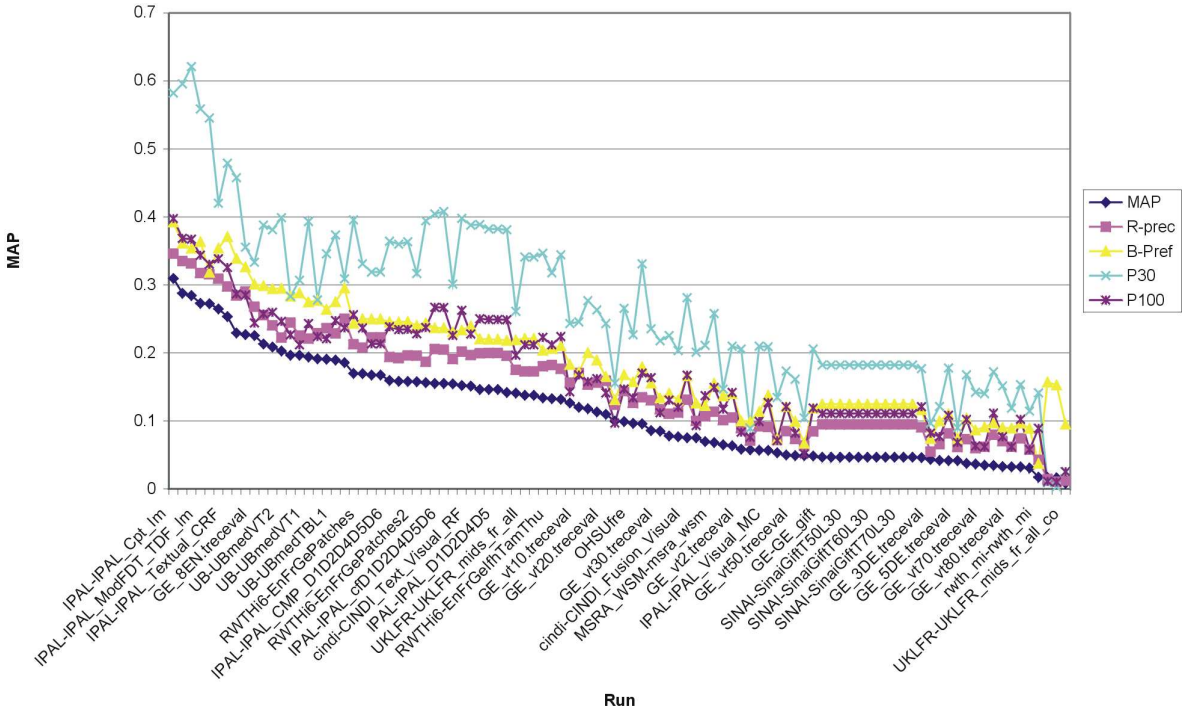



Figure 5: Evaluation results for the best runs of each system in each category, ordered by MAP.

Table 1: Overview of the automatic runs.

| Run identifier | visual | textual | MAP | R–Prec |
|---|---|---|---|---|
| IPAL-IPAL_Cpt_Im | x | x | 0.3095 | 0.3459 |
| IPAL-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-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-CINDI_Fusion_Visual | x | | 0.0753 | 0.1311 |
| MSRA_WSM-msra_wsm | x | | 0.0681 | 0.1136 |
| IPAL-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-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-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-IPAL_Textual_CRF | | x | 0.2534 | 0.2976 |
| OHSU-OHSU_m1 | x | x | 0.1563 | 0.187 |
| cindi-CINDI_Text_Visual_RF | x | x | 0.1513 | 0.1969 |
| cindi-CINDI_Visual_RF | x | | 0.0957 | 0.1347 |

#### 2.7.1 Interactive retrieval

Table 3 shows the submitted interactive runs. The First run shows a very good performance but is still not better than the best automatic run of the same group. The small number of interactive runs is unfortunate but is happening in other evaluation campaigns as well.

### 2.8 Conclusions

The best overall run by the IPAL institute is an automatic run using visual and textual features. In total we can say that interactive and manual runs do not manage to be better than the automatic runs. This may partly due to the fact that most groups submitted many more automatic runs than other runs. This seems 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 but fail to help for the semantic features. Text is very good 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 go a first step toward using automatically annotated images in a multi-modal retrieval system [12]. This year's medical automatic annotation task builds on top of last year: 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 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 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.

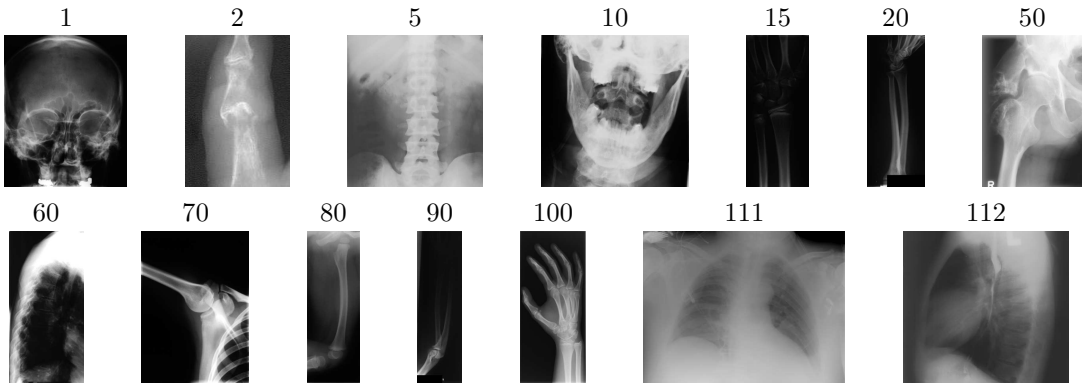



Figure 6: Example images from the IRMA database together with their class numbers

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 [10]. The IRMA code is a multi-axial code for the annotation of medical images. 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 given 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, and 6 classes have only 1‰ or less.

## 3.2 Participating Groups & Methods

In total, 27 groups registered and 12 of these submitted runs. Here for each group a very short 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 PCA transformation. For four of the runs, a support vector machine 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.

**DEU.** The from the Department of Computer Engineering 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 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 use local image descriptors only. Features are dimensionality reduced by a PCA transformation and those runs which use the same features differ in the PCA coefficients kept. The local and 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 feature vector when local and global features are combined. For classification a support vector machine with RBF kernel is used. The group expects the run 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, support vector machines are 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 32x32 pixels and classified using a support vector machine. In the second stage, those decisions where the support vector machine was not sure about the decision, the decision was refined using a classifier that was trained on a subset of the training images. Furthermore, 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 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 GLCM features. For classification multilayer perceptrons 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 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 neighbor classifier, one of the other runs uses support vector machines, and the other uses a maximum entropy classifier. The group expects the run `SHME` to be their best submission.

**RWTHmi.** The IRMA group of the Institute for Medical Informatics Division of the RWTH Aachen University Hospital in Aachen, Germany submitted one run which uses 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 Castelli et al. based on fractal concepts. For classification a nearest neighbor classifier was used. Weights for these features were optimized on the development set.

**UFR.** The Pattern Recognition and Image Processing group from the University 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 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 4-D 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 Systems and Modeling group of the Institute Montefiore from 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 by 20 extremely randomized trees is automatically built based on size-normalized versions of the sub-windows, and operating directly on their pixel values to predict classes of sub-windows. Given this sub-window classifier a new image is classified by classifying sub-windows and combining the classification decisions. 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 submitted one run. The images are scaled to 16×16 pixels and their dimensionality is reduced by PCA transformation. Then a weighted k-nearest neighbor algorithm is applied for classification.

### 3.3 Results

The results from the evaluation are given in Table 4. 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 neighbor classifier using Euclidean distance to compare the images scaled to 32×32 pixels [8]. This classifier yields an error rate of 32.1%. The average confusion matrix of all submitted runs is given in Figure 7. It can clearly be seen that a diagonal structure is reached and thus that on the average many images were classified correctly, but it can also be seen that some classes have high inter-class similarity: in particular, the classes 108 to 111 are often confused and in total many images from other classes were classified to be from class 111, which is the class with the highest amount of training data. Obviously, not all classes are 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; which was to be expected and has been reported in the literature already earlier.

### 3.4 Discussion

The most interesting observation of this years 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 outperform this result this year, 9 of these use support vector machines 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 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.

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 $c$ is the class and $x$ the observation. Unlike last years results, where this technique could not improve the results, clear improvements are possible combining several classifiers [9]: 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. Note, that here no additional parameters were tuned but the classifiers were combined weighted equally.

## 4 Overall Conclusions

The task of the medical automatic annotation task and the non-medical automatic annotation tasks are very similar, but differ in some critical aspects:

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 | rwth-mi | 21.5 |
| * | 13 | CINDI | cindi-svm-sum | 24.1 |
| | 14 | CINDI | cindi-svm-product | 24.8 |
| | 15 | CINDI | cindi-svm-ehd | 25.5 |
| | 16 | CINDI | cindi-fusion-KNN9 | 25.6 |
| | 17 | CINDI | cindi-svm-max | 26.1 |
| * | 18 | OHSU | OHSU-iconGLCM2-tr | 26.3 |
| | 19 | OHSU | OHSU-iconGLCM2-tr-de | 26.4 |
| | 20 | NCTU | dblab-nctu-dblab2 | 26.7 |
| | 21 | MU | I2R-refine-SVM | 28.0 |
| | 22 | OHSU | OHSU-iconHistGLCM2-t | 28.1 |
| * | 23 | ULG | SYSMOD-RANDOM-SUBWINDOWS-EX | 29.0 |
| | 24 | DEU | DEU-3NN-EDGE | 29.5 |
| | 25 | OHSU | OHSU-iconHist-tr-dev | 30.8 |
| | 26 | UTD | UTD | 31.7 |
| | 27 | ULG | SYSMOD-RANDOM-SUBWINDOWS-24 | 34.1 |

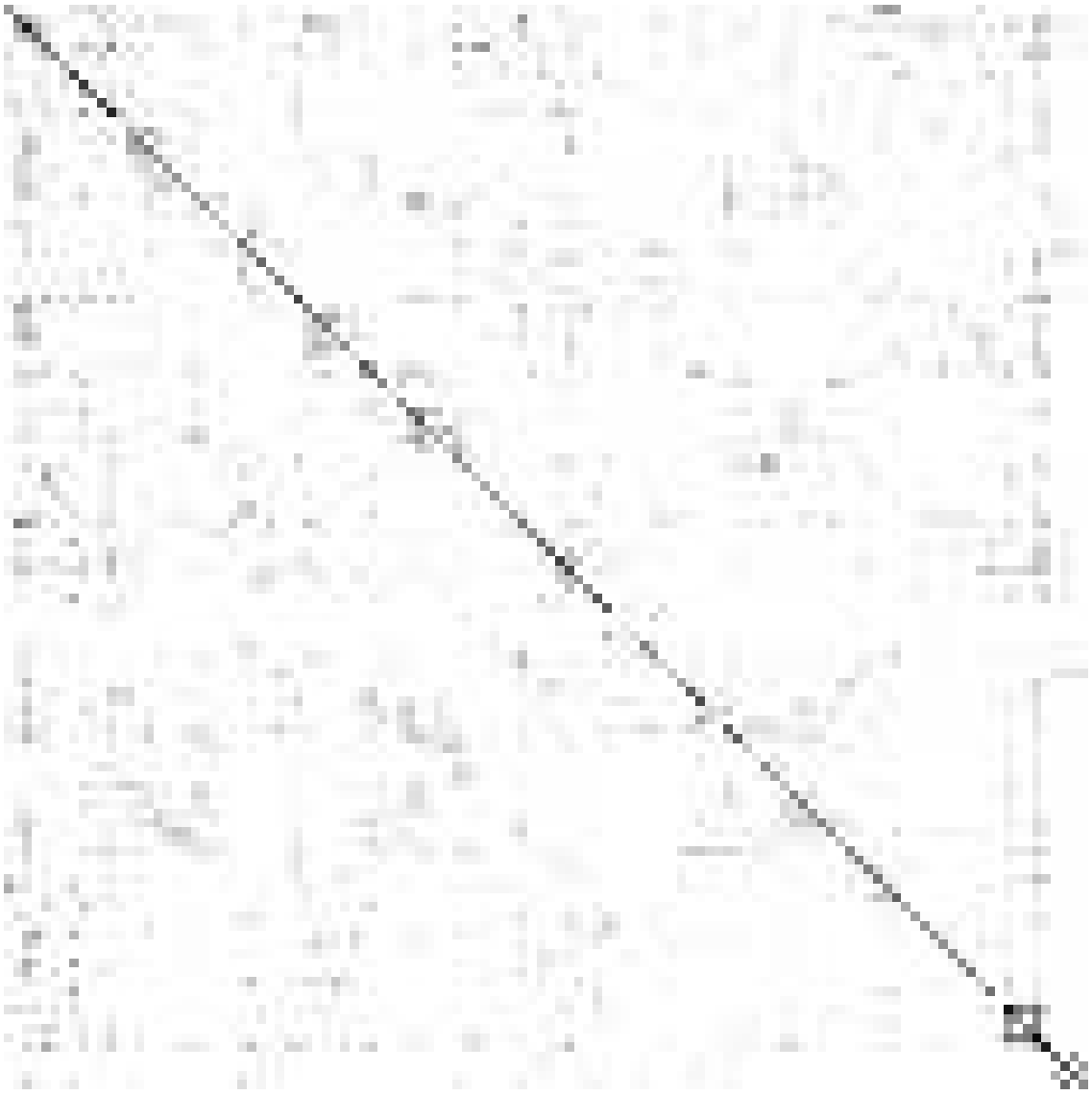


Figure 7: Average confusion matrix over all runs of the medical automatic annotation task. Dark points denote high entries, white points denote zero. On the x-axis, the correct class is given and on the y-axis the class to which images have been classified is given. For visualization purposes values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

- Both tasks provide a relatively large training set and a disjunct test set. Thus, in both cases it is possible to learn a relatively reliable model for the training data (this is somewhat proven for the medical annotation task, and below we also show this for the non-medical task..
- Both tasks are multi-class/one object per image classification tasks. Here they differ from the PASCAL visual classes challenge which addresses a set of object vs. non object tasks where several objects (of equal or unequal type) may be contained in an image.
- The medical annotation task has only gray-scale images, whereas the non-medical tasks has mainly color images. This is probably most relevant for the selection of descriptors.
- The images from the test and the training set are from the same distribution for the medical task, whereas for the non-medical task, the training images are rather clutter-free and the test images contain a significant amount of clutter. This is probably relevant and should be addressed when developing methods for the non-medical task. Unfortunately, our models currently do not address this issue which probably has a significant impact on the results.

## 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.

[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.

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

[6] 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.

[7] 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.

[8] 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.

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

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.