Hes·so// VALAIS WALLIS

Business Information
Systems

Evaluation and image retrieval

Henning Müller

Thomas Deselaers

UNIVERSITÉ DE GENÈVE

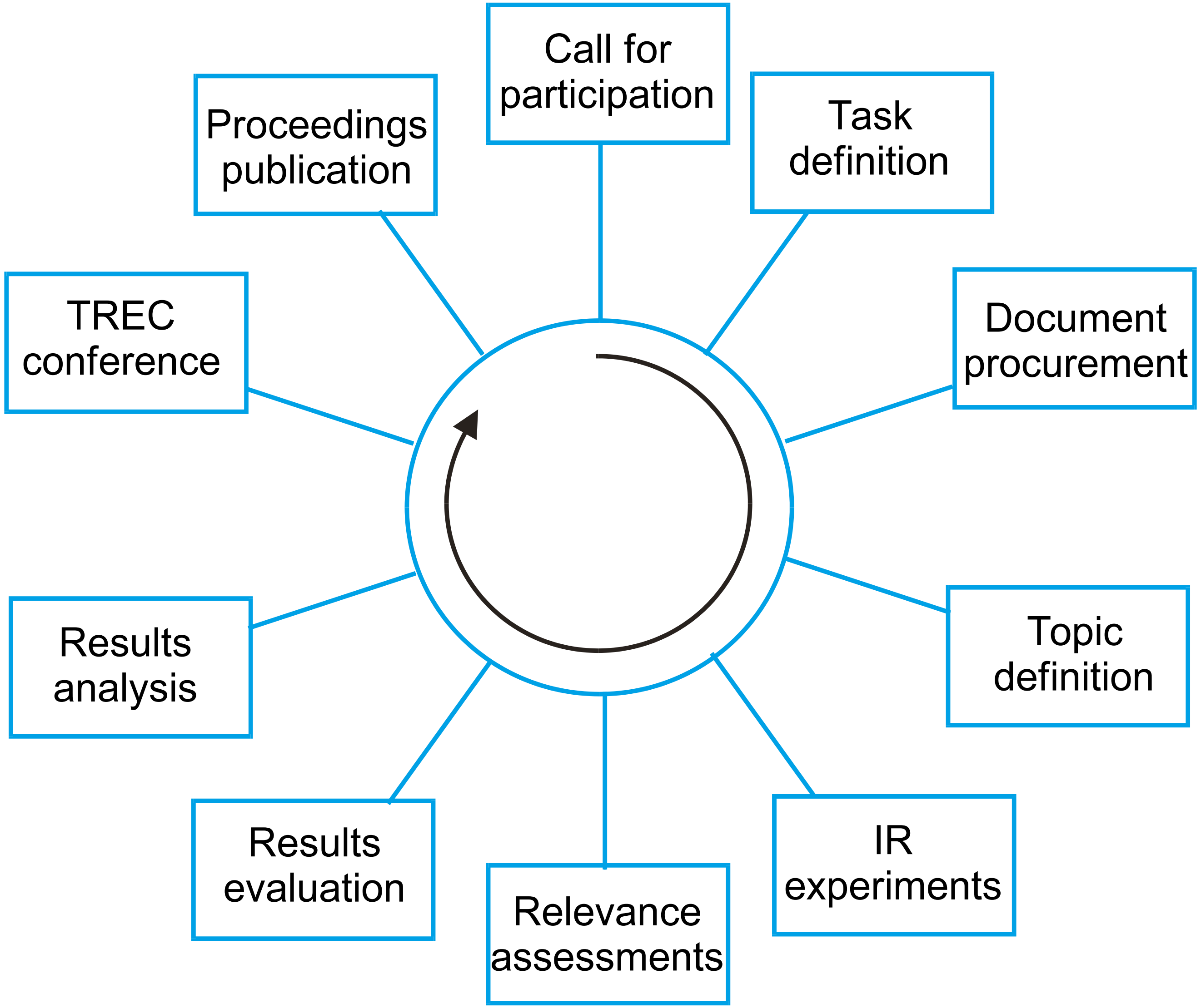
Overview

- Information retrieval evaluation
  - TREC
- Multimedia retrieval evaluation
  - TRECVID, ImageEval, Benchathlon, …
- ImageCLEF
  - Past
  - Future

Information retrieval evaluation

- Started very early (1960s, in part as a theoretical discipline …)
  - Cranfield tests, Smart
- TREC became a role model for benchmarks with many spin-offs (TRECVID, CLEF, …)
  - Yearly circle of events
  - Relevance-based evaluations, …
  - Mainly system-oriented evaluation
- Still, much can be criticized
  - Measures, interactive retrieval, …

A yearly circle



Visual retrieval evaluation

- Little systematic evaluation in first years of research (1990-2000)
  - Some papers on methodologies
  - Benchathlon to foster discussions
- Since then, evaluation has come a long way …
- TRECVID, ImageCLEF, INEX MM, ImageEval, …
  - Improvement in performance can be shown
  - Techniques can be compared
- Methodologies and user models can be criticized
  - Not all research can be benchmarked
  - Innovation instead of pure performance

Axes for benchmarks

- Databases
- Tasks/topics
  - Including experts for relevance judgements
- Participants
  - Techniques to compare
- Ground truth, gold standard
- Performance measures

Problems of IR benchmarks

- Funding
- Access to visual datasets
- Motivate participation (everybody is afraid to loose)
- Partners from industry
- Realistic tasks and user models
- Ground truthing (costly, ambiguous)
- Organisational issues
- Proving advances and benefits

CLEF - ImageCLEF

- Cross Language Evaluation Forum
  - Started as track in TREC (Text Retrieval Conference,1997)
- Independent workshop since 2000
- Multilingual information retrieval
  - Collections are multilingual
  - Queries are in a language different from the collection
- Good framework, registration, legal issues, proceedings in Springer LNCS, …

History of ImageCLEF

- 2003: first image retrieval task, 4 participants
- 2004: 17 participants for three tasks (~200 runs)
  - Medical task for visual image retrieval added
- 2005: 24 participants for fours tasks (~300 runs)
  - Two medical tasks
- 2006: 30 participants for four tasks (~300 runs)
  - LTU database of objects for object classification
- 2007: 35 participants (>1000 runs)
  - Hierarchical classification
- 2008: 45 participants submitted results (>2000 runs)
  - 63 registrations, wiki task

ImageCLEF 2008

- ImageCLEF/Quaero workshop on image retrieval evaluation
  - To motivate visual retrieval community
- Ad-hoc retrieval with query in a different language
  - Photo collection, vacation pictures of an agency
- Concept detection task
- Medical Retrieval task
  - Collection of ~70’000 images with annotations
- Medical classification task
  - Hierarchical classification
- Wikipedia retrieval task
- Interactive retrieval (using a FlickR API)

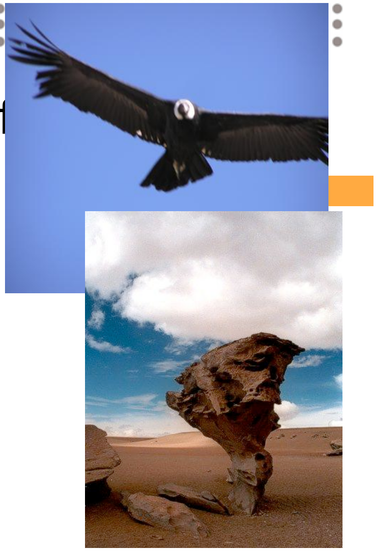
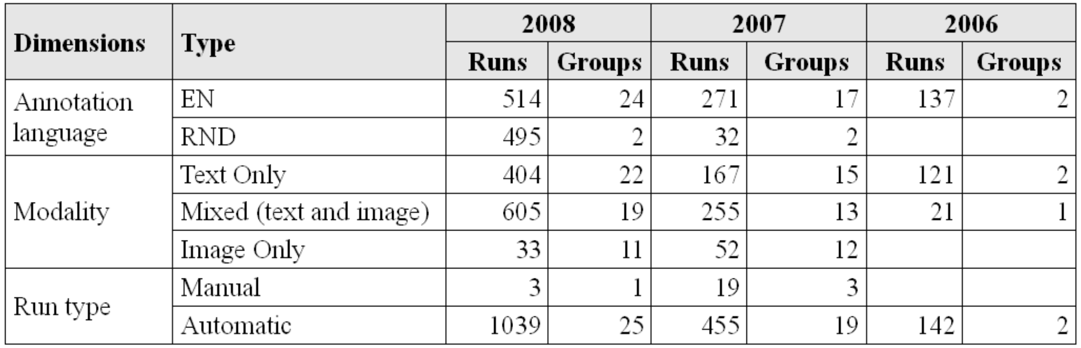


Photo retrieval 2008

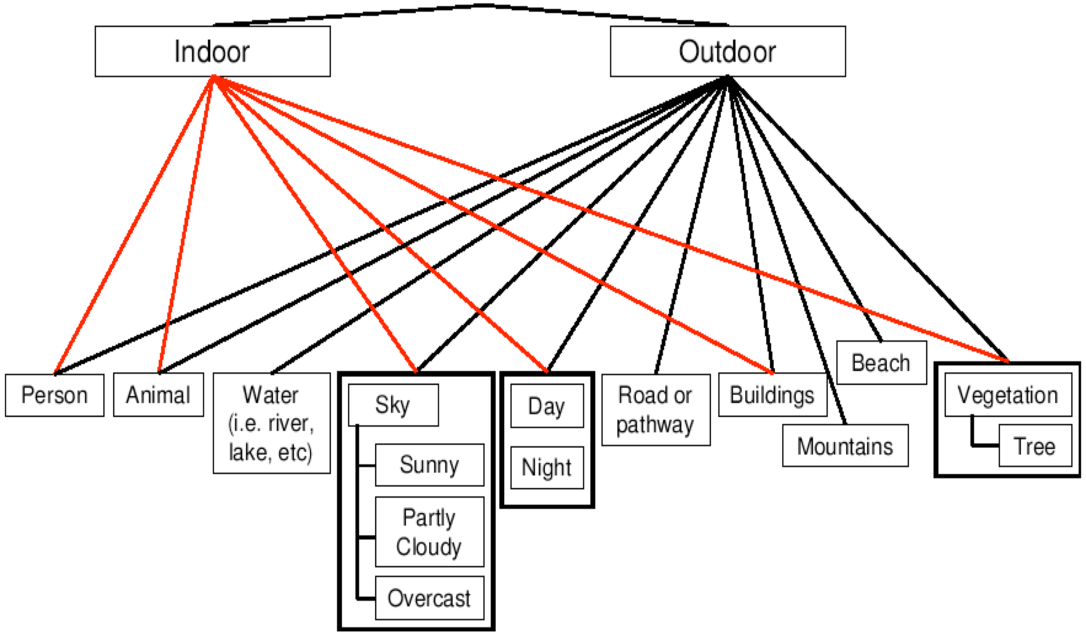
- Promote diversity in retrieval
  - Evaluated using Cluster Recall
- Very strong participation
  - Most participants used two stage process: perform ad-hoc retrieval; then cluster results
- Analysis of results showed
  - Standard retrieval does not promote diversity
  - Choice of language negligible for results
  - Combining content and concept-based methods gives best results



Dimensions	Type	2008		2007		2006	
		Runs	Groups	Runs	Groups	Runs	Groups
Annotation language	EN	514	24	271	17	137	2
	RND	495	2	32	2		
Modality	Text Only	404	22	167	15	121	2
	Mixed (text and image)	605	19	255	13	21	1
	Image Only	33	11	52	12		
Run type	Manual	3	1	19	3		
	Automatic	1039	25	455	19	142	2

Visual Concept detection

- Small hierarchy of concepts for annotation
- Purely visual concept detection works well
- Local features such as SIFT outperform other techniques
- Link with photo retrieval, but only used by a single group



WikipediaMM retrieval task

Business Information Systems

- Semi-Structured annotation together with images
  - This year annotation and topics in English
- Not all topics contained images
  - Bias against visual retrieval
- Text retrieval works well
  - Visual concepts can improve overall performance
- Participants are judges

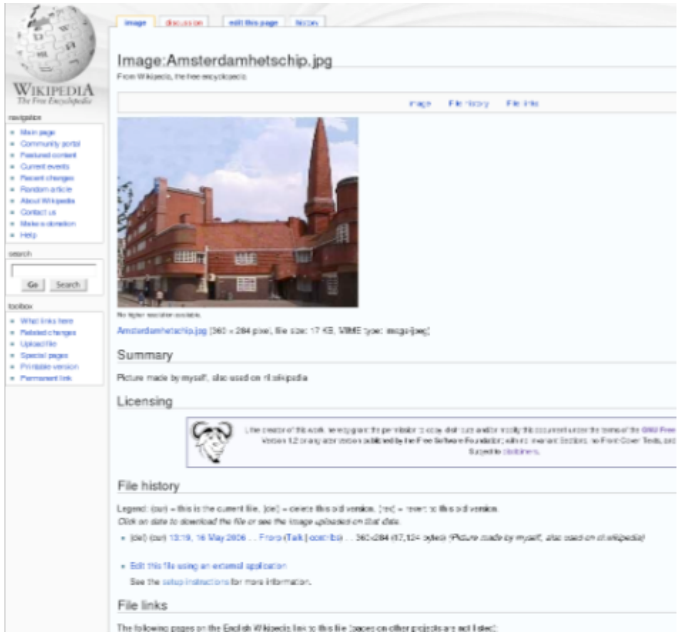


Image:Amsterdamhetschip.jpg

From Wikipedia, the free encyclopedia

No higher resolution available.

Amsterdamhetschip.jpg (360 × 284 pixel, file size: 17 KB, MIME type: image/jpeg)

Summary

Picture made by myself, also used on nl.wikipedia

Licensing

I, the creator of this work, hereby grant the permission to copy, distribute and/or modify this document under the terms of the GNU Free
Version 1.2 or any later version published by the Free Software Foundation; with no invariant Sections, no Front-Cover Texts, and
Subject to disclaimers.

File history

Legend: (cur) = this is the current file, (del) = delete this old version, (rev) = revert to this old version.

Click on date to download the file or see the image uploaded on that date.

- (del) (cur) 13:19, 16 May 2006 . . Foroa (Talk | contribs) . . 360×284 (17,124 bytes) *(Picture made by myself, also used on nl.wikipedia)*

- Edit this file using an external application

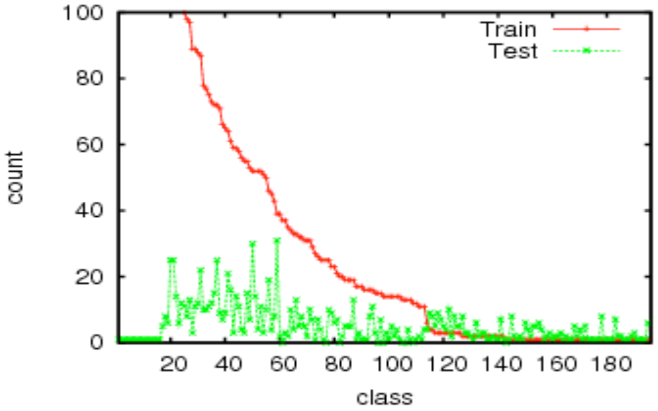
See the setup instructions for more information.

File links

The following pages on the English Wikipedia link to this file (pages on other projects are not listed):

Medical annotation task 2008

- Again a hierarchy of classes for visual classification
  - Distribution of classes in training and test data not equal
  - Forced to use confidence on a hierarchy level



- Local features outperform global ones
- Machine learning techniques are key to success
- Results of past years published in special issue

Tasks for the medical task

- Realistic!!
  - – Based on independent expert opinions
  - – Based on surveys (Portland, Geneva)
  - – Based on log files (health on the net media search, medline)
- Retrieval with varying degree of visualness
  - – A little subjective
- Afterwards analysis of results per task
  - – Analyze ambiguity for judges (double judgments)
    - Kappa analysis

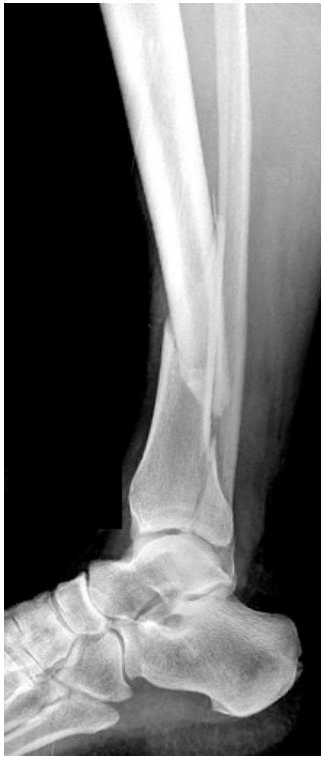
Task examples

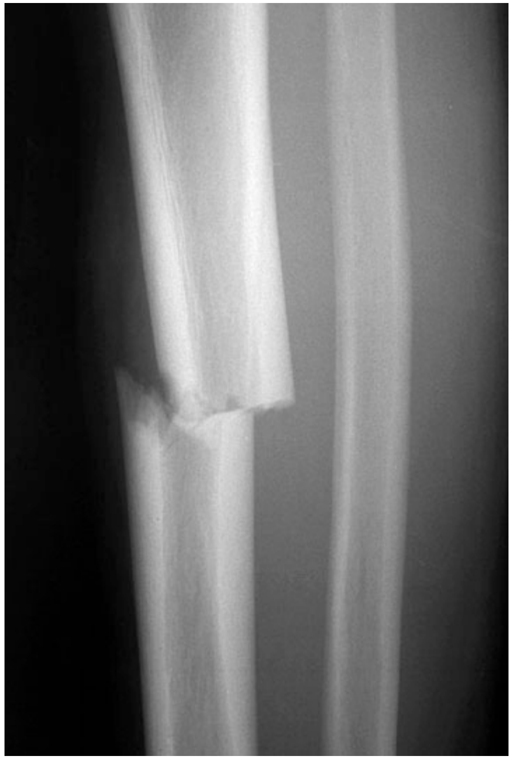
1.4

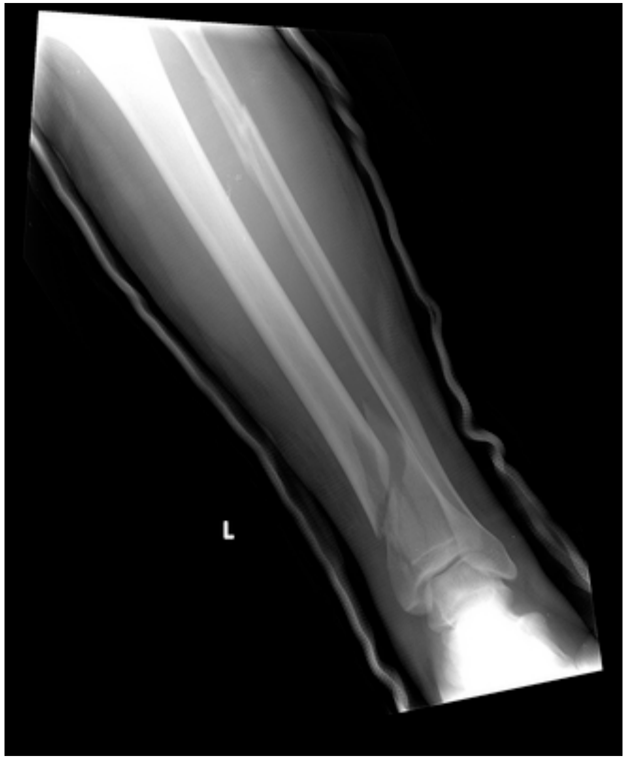
Show me x-ray images of a tibia with a fracture.

Zeige mir Röntgenbilder einer gebrochenen Tibia.

Montre-moi des radiographies du tibia avec fracture.







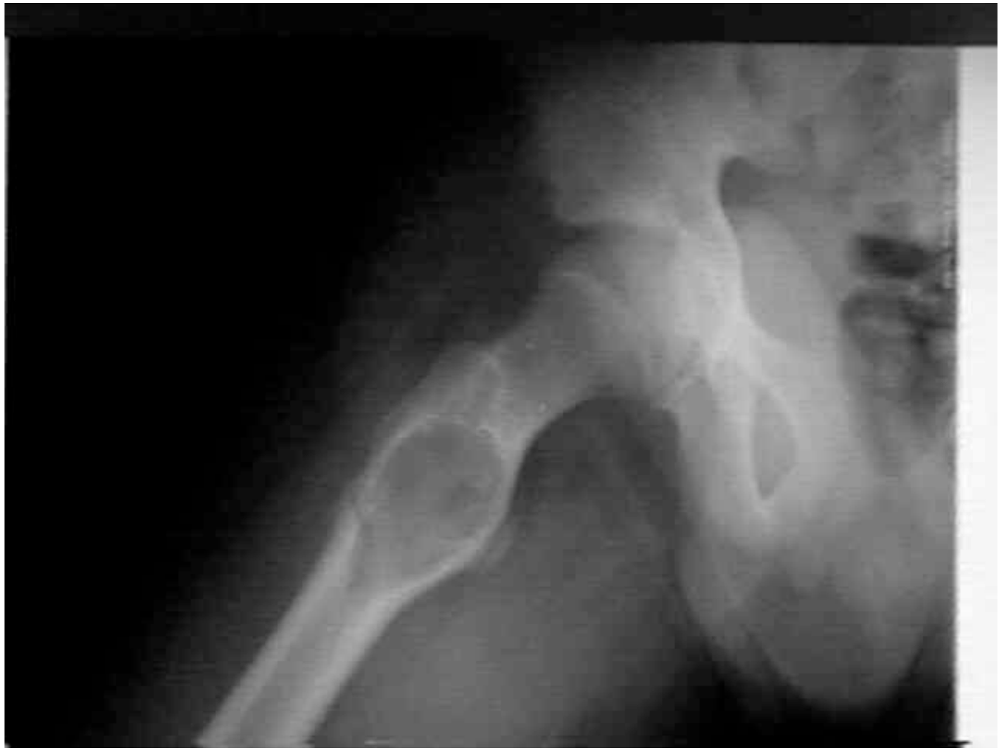
Task examples

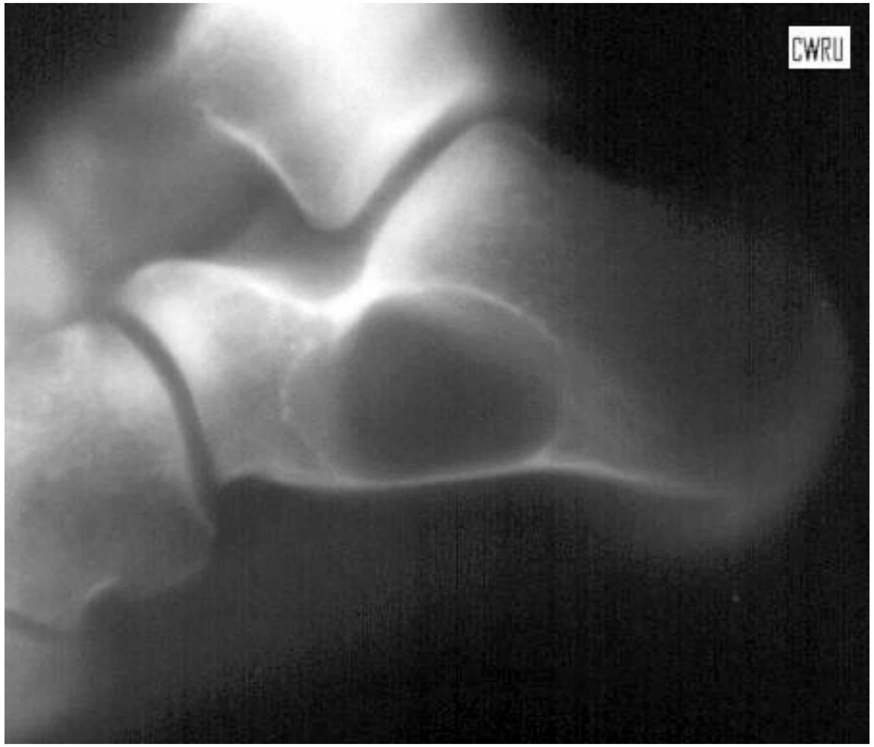
3.6

Show me x-ray images of bone cysts.

Zeige mir Röntgenbilder von Knochenzysten.

Montre-moi des radiographies de kystes d'os.





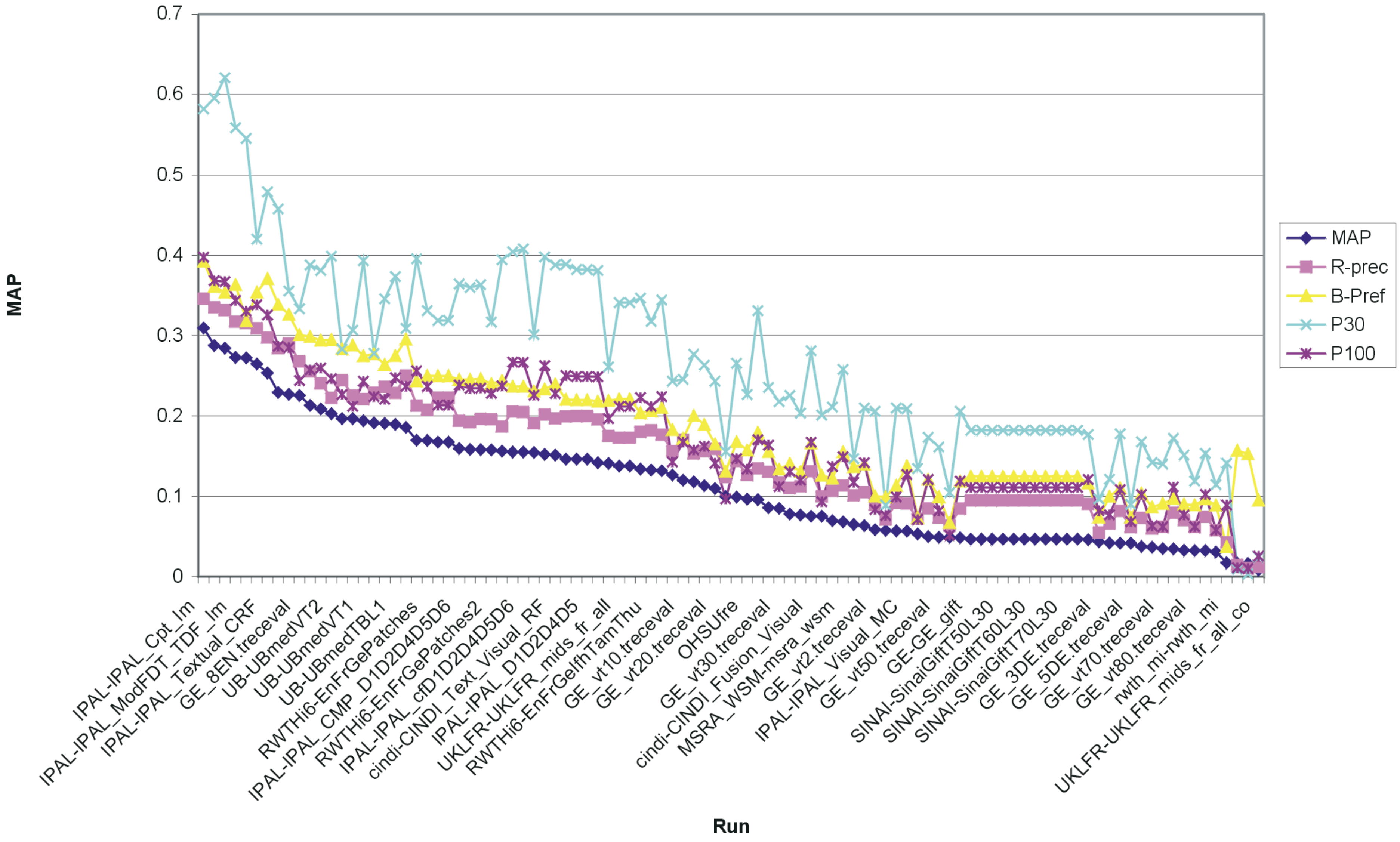
Ground truthing

- Retrieval
  - Expensive task with real users!
    - Funding from NSF, help from participants
  - Pooling is used with varying number depending on submissions
  - Judgment scheme: relevant – partially – non-relevant
    - Describe all categories exactly!!
  - Double judgments to analyze ambiguity
    - Good systems stay good with any judge
- Interactive
  - Participants evaluate themselves (time, Nrel)

Evaluation

- Categories for media used
  - Visual, textual, mixed
- Categories for interaction used
  - Automatic, feedback, manual modification
- Still: Mean Average Precision as a lead measure
  - Correlates very well with other measures
  - BPref, P(10-50) used for comparison
- Many ideas on how to find better measures
  - No resources to pursue this

MAP and other measures

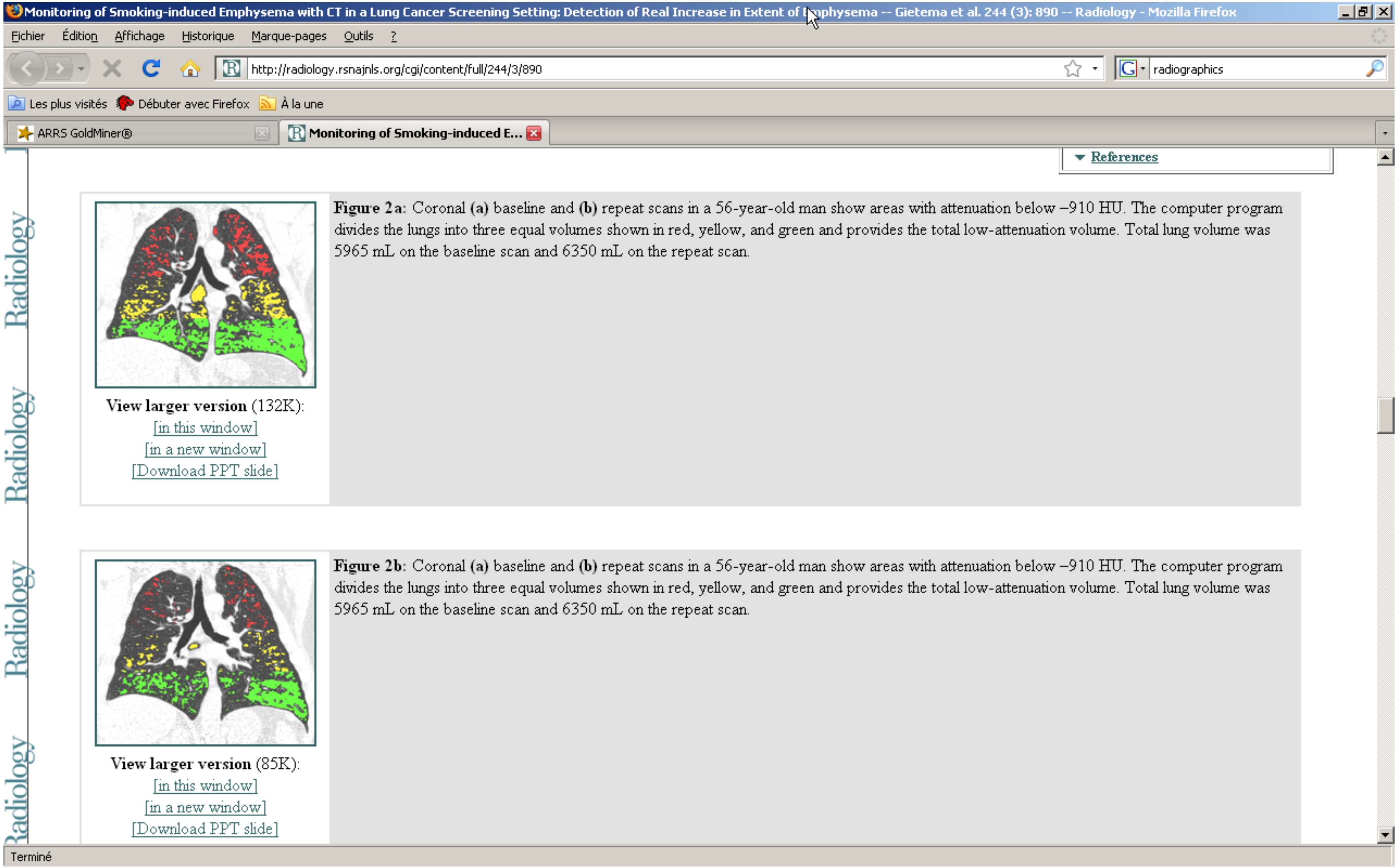


Workshop

- Event for discussions among participants
  - Mix visual and text retrieval communities
  - Learn from results of others
- Oral presentations are selected based on novelty of techniques not on performance
- Every participant can present a poster
- Presentation of the main findings

- Feedback is very positive and participants do not regret their participation

Example from the database 2008



Monitoring of Smoking-induced Emphysema with CT in a Lung Cancer Screening Setting: Detection of Real Increase in Extent of Emphysema -- Gietema et al. 244 (3): 890 -- Radiology - Mozilla Firefox

http://radiology.rsnajnls.org/cgi/content/full/244/3/890

Figure 2a: Coronal **(a)** baseline and **(b)** repeat scans in a 56-year-old man show areas with attenuation below −910 HU. The computer program divides the lungs into three equal volumes shown in red, yellow, and green and provides the total low-attenuation volume. Total lung volume was 5965 mL on the baseline scan and 6350 mL on the repeat scan.

View larger version (132K):
[in this window]
[in a new window]
[Download PPT slide]

Figure 2b: Coronal **(a)** baseline and **(b)** repeat scans in a 56-year-old man show areas with attenuation below −910 HU. The computer program divides the lungs into three equal volumes shown in red, yellow, and green and provides the total low-attenuation volume. Total lung volume was 5965 mL on the baseline scan and 6350 mL on the repeat scan.

View larger version (85K):
[in this window]
[in a new window]
[Download PPT slide]

ImageCLEFmed 2008

- Images and full-text articles of Radiology/ Radiographics (thanks to the RSNA!)
  - Captions of the figures with detailed information on the figures, subfigures
  - The kind of data that clinicians search
- Detailed search tasks may not be the most common for diagnosis, rather teaching
- More adapted for text retrieval, image analysis has to be done with care

Some results

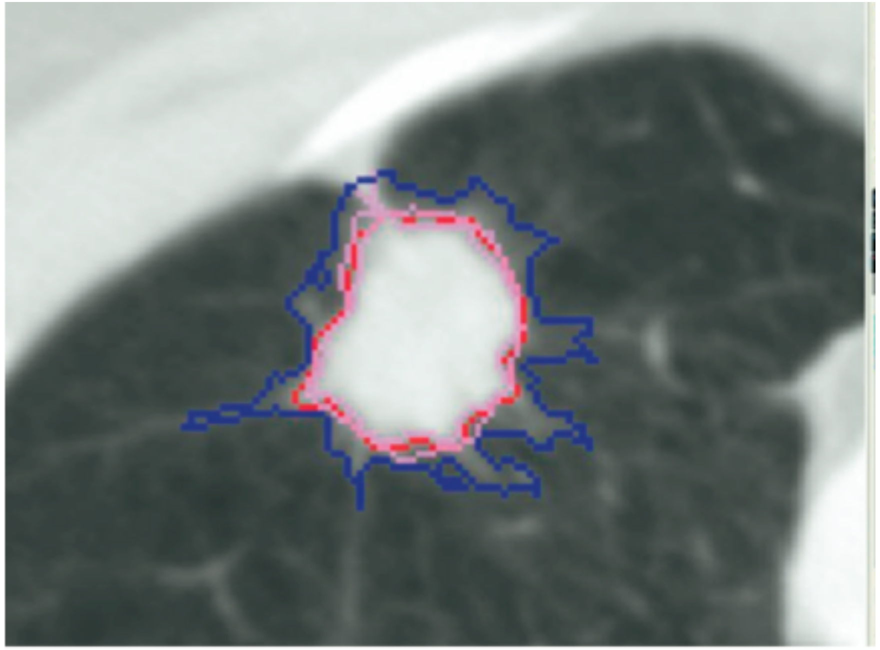
- Visual retrieval has often good early precision but poor recall
- Visual features can be useful for specific queries
  - This can be detected more or less automatically
- Multimodal retrieval has most potential
- Visual classification has improved significantly
- Relevance feedback and interactive retrieval are rarely used
  - (lack of manpower, non-interactive setups)

ImageCLEFmed 2009

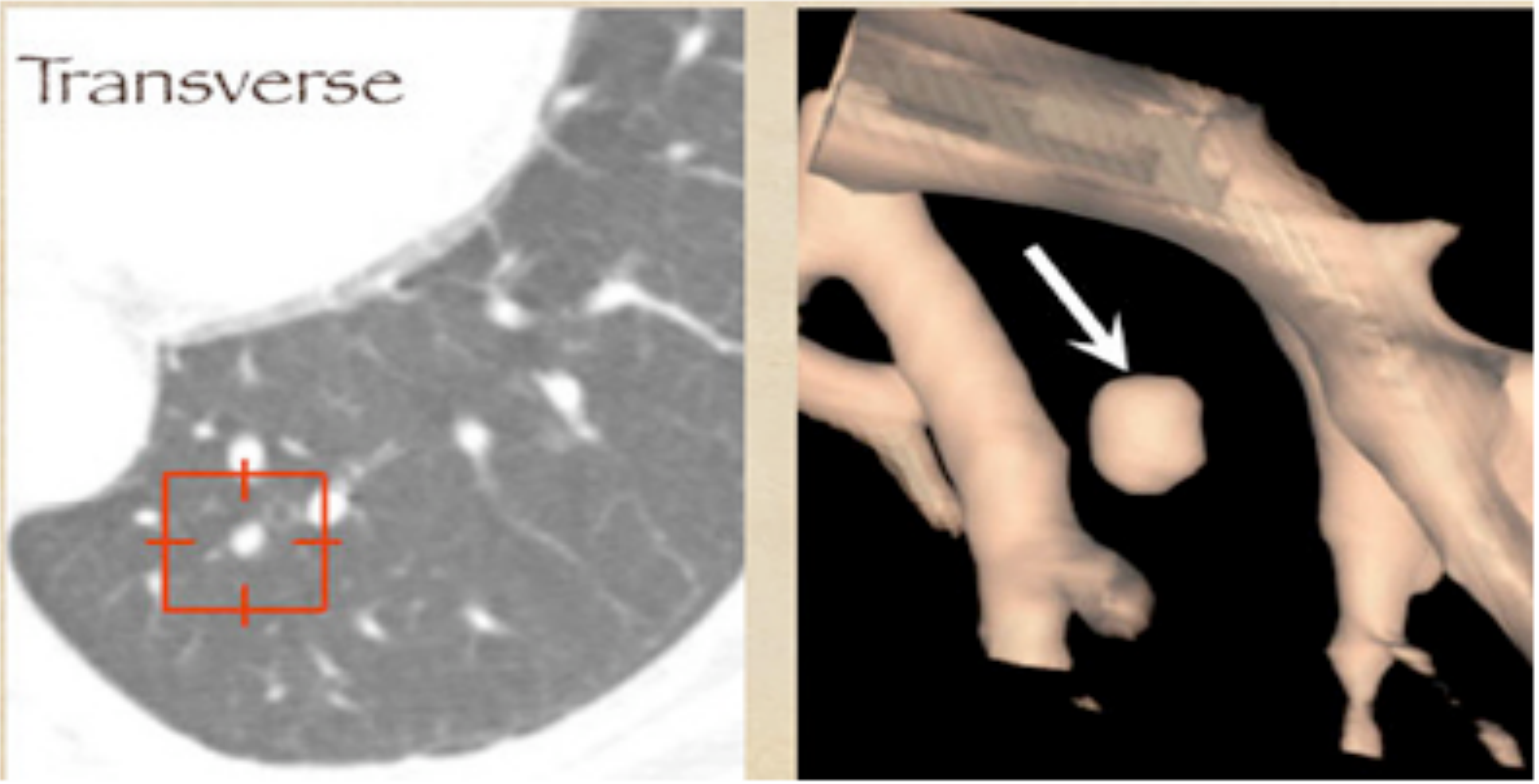
- Search for similar cases in the literature
  - Several sorts of images (xray, CT, MRI)
  - Use incomplete data (no textual information on modality, pathology)
  - Much more realistic scenario! Clinician in the process of solving a difficult case
- Hard task: text processing might not work
  - Fusion of very varied data is an important topic

ImageCLEF 2009 medical classification

- Nodule detection in lung CT images
- Image database from the LIDC
  - Supplied with the help of the NCI
- Small region of interest to detect
  - Exact place and size
  - Potential 3D task



3D task



CT finding (left) has the appearance of an adjacent vessel in transverse-section reconstruction and was not called by any of the four LIDC readers. After viewing transverse, coronal, sagittal, and volume-rendered reconstructions (right), all four university readers called the finding a lung nodule.

ImageCLEF 2009

- Robot Vision task
  - – Details to be defined
- Again a wikipedia task
- Maybe a new photo retrieval task
  - – Larger dataset
  - – Maybe in connection with an ontology-based annotation (Theseus project)

Conclusions

- Evaluation is important to be able to compare techniques
  - Performance alone is not the only goal
- Benchmarks provide a basis usable for many researchers
  - Avoiding much double work of creating data
  - They are hard to organise and harder to fund
- ImageCLEF has had an impact on data/techniques used by participant in visual retrieval

References

- IR
  - G. Salton, The SMART Retrieval System, Experiments in Automatic Document Processing, Prentice Hall, Englewood Cliffs, New Jersey, USA, 1971.
  - C. W. Cleverdon, Report on the testing and analysis of an investigation into the comparative efficiency of indexing systems, Technical report, Aslib Cranfield Research Project, Cranfield, USA, September 1962.
  - E. M. Voorhees and D. Harmann, Overview of the seventh Text {RE}trieval Conference (TREC-7), The Seventh Text Retrieval Conference, pages 1--23, Gaithersburg, MD, USA, November 1998.
- Visual retrieval
  - Alan F. Smeaton, Paul Over, Wessel Kraaij, Evaluation campaigns and TRECVid, Proceedings of the 8th ACM international workshop on Multimedia information retrieval table of contents, pages 321-330, Santa Barbara, California, USA, 2006.
  - N. J. Gunther and G. Beretta, A benchmark for image retrieval using distributed systems over the internet: BIRDS-I, Technical report, HP Labs, Palo Alto, Technical Report HPL-2000-162, San Jose, 2001.
  - T. Deselaers, T. Deserno, H. Müller, Automatic Medical Image Annotation in ImageCLEF 2007: Overview, Results, Discussion, Pattern Recognition Letters, volume 29, pages 1988-1995, 2008.

- P. Clough, H. Müller, T. Deselaers, M. Grubinger, T. Lehmann, J. Jensen, W. Hersh, The CLEF 2005 Cross-Language Image Retrieval Track, Springer Lecture Notes in Computer Science LNCS 4022, pages 535-557, 2005.
- H. Müller, S. Marchand-Maillet, T. Pun, The Truth About Corel - Evaluation in Image Retrieval, In Proceedings of The Challenge of Image and Video Retrieval (CIVR2002), Springer Lecture Notes in Computer Science LNCS2383, pages 38-49, London, UK, July 2002.
- H. Müller, W. Müller, D. McG. Squire, S. Marchand-Maillet, T. Pun, Performance Evaluation in Content-Based Image Retrieval: Overview and Proposals, Pattern Recognition Letters (Special Issue on Image and Video Indexing), 22, 5, pages 593-601, 2001. H. Bunke and X. Jiang Eds.
- H. Müller, J. Kalpathy-Cramer, C. E. Kahn Jr., W. Hatt, S. Bedrick, W. Hersh, Overview of the ImageCLEFmed 2008 Medical Image Retrieval Task, Springer Lecture Notes in Computer Science, 2009.
- W. Hersh, H. Müller, J. Kalpathy-Cramer, E. Kim, X. Zhou, The consolidated ImageCLEFmed Medical Image Retrieval Task Test Collection, Journal of Digital Imaging, 2008.