Tutorial on Medical Image Retrieval - evaluation -

Medical Informatics Europe
28.08.2005

Henning Müller, Thomas Deselaers

Service of Medical Informatics
Geneva University & Hospitals, Switzerland
Aachen Technical University, Germany

Overview

- Introduction to (image retrieval) evaluation
- Parts of an image retrieval benchmark
  - Data sets
  - Query tasks and topics
  - Ground truth
  - Evaluation measures
  - Benchmarking events
- Current initiatives
  - TRECVID
  - Benchathlon
  - imageCLEF
- Conclusions

Introduction

- What do we actually want to evaluate?
  - Realistic scenarios
  - Real user needs
  - What can we do if it is not used in practice?
- Text retrieval has a long experience in evaluation
  - Cranfield (early 60s), Smart, TREC, CLEF
  - What can we use and what not?
  - More commercial interest
  - First systems in 1960s were more theoretical
- Usability testing as well?

Usability testing, human factors

- Tests how real users operate with the system
  - User interface
  - Easy and quick to use
  - Adherence to interface standards
  - Novice vs. Advanced user mode
- Interactivity tests
  - Speed is important
- Result needs to be explained to the user
  - On screen feedback

Evaluating clinical information systems

- Validation of algorithms on test data
- Evaluation of the results on real data sets
- Clinical impact
  - Through user tests, improved diagnoses
- Outcome, does the use reduce the patient length of stay or the reduced use of system resources
  - Often hard to prove

- We are still in an extremely early stage for image retrieval

The need for evaluation

- Without evaluation there is no proof of performance
  - No improvement can be shown
  - Techniques cannot be compared
  - Techniques will not have any commercial success
    - We need to see how far image retrieval has come with respect to this, can we answer real user needs?
- Systematic evaluation can bring big improvements and deliver important results
  - Cranfield tests
  - TREC
  - Other domains
    - Compression, segmentation, watermarking of images, …

History of image retrieval evaluation

- Example results of one query, then several queries
- Use of databases (Corel, Vistex) containing very similar images
  - Problem: different subsets
- Use of self-defined measures
  - Show clustering, often only one measure
  - Definition of invariant measures (generality, invariant PR graphs)
- Use of standardized measures
  - Recall, precision, normalized average precision (MPEG7), mean average precision (TREC)
- Why is it so hard to compare any two retrieval systems on the same basis?

Parts needed for a benchmark

- Data sets
  - Corel, Washington, Benchathlon, MPEG-7, Casimage
- Query topics and tasks
  - Definition based on real world tasks is needed!
- Ground truth
  - Implicitly used through Corel categories
  - Otherwise expensive
- Evaluation measures
- Benchmarking events

Image data sets available

- Corel
  - Not sold anymore, but thumbnails possible
- University of Washington
  - Groups of photographs from various regions
- MPEG-7 (copyrighted)
- Benchathlon
  - Images of people
- Casimage (medical images, and multilingual text), MIRC
- Corbis test set (text and images)
  - Which conditions?
- NIH publishes all the created databases but non for retrieval, so far
- Size matters!

Query tasks and topics

- Very few analyses of user behavior are available
  - Journalists queries (Finland)
  - Image archive use (England)
  - Trademark retrieval is fairly well defined
  - Study on medical images at OHSU, HUG in 2005
- How can we define real-world tasks?
  - They will have to be based on the databases available
  - Survey of medical teaching file users
  - Problem: Almost no retrieval systems in routine use
  - How can we find out real behavior without a standard use of the systems?

Ground truth, Gold standard

- Expensive to define
  - Will need to involve real users
  - More than one set is good to model subjectivity
  - Pooling reduces complexity slightly (TREC methodology)
- Classification of images is practical but change of databases might be hard (complete annotation)
- Databases and ground truth will need to be changed from time to time (regularly)
- Community effort would be great
  - Common project (EU, NSF, ?), financing needed
  - Annotation?

Performance measures

- Standards that are easy to interpret exist
  - Precision, recall, norm. average rank (MPEG-7), …
  - Mean average precision to create a ranking at TREC
- One measure is not enough
  - Although measures are strongly correlated
- Normalization of collection size (generality) is not needed
  - Difficulty of query task can be described in other ways
  - Comparison with different databases is not useful
- Measures do not pose a critical problem for evaluation

Performance measures (2)

- Precision

$$P = \frac{\text{number of relevant images retrieved}}{\text{number of all images retrieved}}$$

- Recall

$$R = \frac{\text{number of relevant images retrieved}}{\text{number of all relevant images in the DB}}$$

Benchmarking event

- Needed for content-based visual information retrieval!!!
- A friendly event that should help everyone
  - Such as trec, clef
- Co-located with conferences where people go anyways to reduce costs
  - Benchathlon at SPIE electronic imaging
  - CLEF at ECDL
- Feedback and acceptance from the community is important
  - But how to motivate research groups?
  - Databases, other benefits

A technical infrastructure for evaluation

- Results send in offline
  - TREC, CLEF
- Interactive user evaluations

- Automatic solution based on a standard communication protocol
  - MRML, solutions exist
  - Web-based evaluation procedure allows quick evaluations after an event
  - Harder to get acceptance

TRECVID

- Video retrieval at TREC, now a separate workshop
- Started in 2001
- 12 participants in 2001, 24 in 2003, 33 in 2004
- 130 hours of video in 2001
- Accepted in the community, proceedings have an impact, new tasks added every year
- Financing through TREC, domain seems important and databases are available (news)
- Speech and captions provide important semantic information

TRECVID tasks

- Shot boundary detection
  - Cut or gradual
- Story segmentation
  - One news story, contains several shots
- Feature extraction
  - Concept extraction: indoor, outdoor, speech, people, train, boat, road, Bill Clinton, …
- Search
  - Human information need is expressed in text+ multimedia
  - Results are a ranked shot list

Benchathlon

- Goal was to create a forum for the discussion on evaluation of image retrieval systems and the creation of an evaluation infrastructure
- Situated at SPIE electronic imaging
- Started in 2001, after discussions in 2000 and an outline document on such a benchmark
- 2002: 5 papers
- 2003: 8 papers
- 2004: only discussions among participants
- 2006: special session on evaluation planned

imageCLEF

- Located at Cross Language Evaluation Forum(CLEF)
- Goal is to evaluate the retrieval of images through multi-lingual information retrieval
- 2003: first image retrieval task, 4 participants
  - Queries in different languages than the English collection annotation, image is part of the query
- 2004: 17 participants for three tasks (~200 runs)
  - Medical task for visual image retrieval added where the query topic is an image, only
- 2005: 24 participants for fours tasks (~300 runs)
  - Two medical tasks, one retrieval and one on classification
  - 36 groups inscribed: much interest in the data

imageCLEF methodology

- Based on the TREC/CLEF methodologies
  - Schedule for participation (January to September)
  - Release of data to participants, then query tasks
  - After result submission, pooling and ground truthing
  - Event to compare results
  - Proceedings with an impact
- Still in a learning phase as only in the third year
- New tasks have been added
  - Interactive query/retrieval in 2004
  - Medical classification, visual only in 2005
  - Tasks need to vary every year to cover new grounds

imageCLEF St. Andrews

Pictures of English lighthouses

イングランドにある灯台の写真

Fotos de faros ingleses

Kuvia englantilaisista majakoista

Bilder von englischen Leuchttürmen

Record ID:	JV-044809
Short title:	The Smeaton Tower, Plymouth
Long title:	Plymouth Hoe. The Smeaton [Lighthouse] Tower.
Location:	Devonshire, England
Description:	Red and white striped lighthouse on coastal cliff with harbour and town beyond, and substantial building on cliff terrace below
Date:	Registered 1904
Photographer:	J Valentine & Co
Categories:	[lighthouses ‖ beacons & lighthouses ‖ Devon all views ‖ Collection - J Valentine & Co]
Notes:	JV-44809 pc/mb(or possibly 44810)TECH: Coloured

- 28.000 images, 25 queries
- Submissions include visual and textual runs and a large variety of techniques
- 2005: three example images per query, tasks taking into account visual content

imageCLEFmed

- More than 50.000 images, 25 query tasks
- Goal was to model search for information by medical specialists
- Teaching file databases, tasks chosen based on a user survey, three classes of topics
  - Ground truthing by medical doctors
- Submissions include automatic and manual submissions and several techniques
  - Text only is better than visual only
  - Best systems combine visual and textual (MAP 0.28)

Image classification

- IRMA group as organisers: 9000 training images from 57 classes, 1000 evaluation images
  - All images are annotated in IRMA code
- Visual properties only, but annotation is available in German and English
  - Goal: keep a challenging task for the visual community
- 12 participants, variety of techniques for features and classification
- Best results: 87.4% correct

imageCLEF results

- Visual classification proves popular
  - More time needed to develop optimized algorithms
- Textual is on average better than visual retrieval
  - Important to have various semantic levels of queries
- Best results were obtained by combining textual and visual features
  - Dependent on the features, though
- Most groups wanted test data, which was not really available this year
  - 2004 tasks were very different

Conclusions

- Evaluation is essential for any research domain to prove the system performance
- Benchmarking events advance science and everybody profits
- Data sets and feedback from real users is crucial for future tasks
  - More studies on this are needed
  - Data sets need to be made available for all articles published if possible