
# The Cross Language Image Retrieval Track: ImageCLEF

## Breakout session discussion


# Overview

- General comments/feedback
  - about the 2006 event
- Proposals for 2007
  - Photographic image retrieval task
  - Medical image retrieval task
  - Medical image annotation and object classification tasks
- Participants feedback


# General comments/feedback

- Tasks offered
  - are they realistic/useful/beneficial?
  - what are the use cases?
- Collections used
- Topics and assessments
- Evaluation measures
- Organisation of the tasks
- Information provided about results
- Provision of resources
  - sharing resources between participants
- Interactive experiments


# Photographic Retrieval Task

## *ImageCLEFphoto*

# Feedback from survey (2006)
*Selected results*

- Document Collection and Query Topics

| | strongly disagree | disagree | neutral | agree | strongly agree |
|---|---|---|---|---|---|
| The IAPR TC-12 Benchmark is an appropriate collection for ImageCLEFphoto. | | | | 5 | 2 |
| The database represents a realistic set of real-life still natural images. | | | | 7 | |
| The quality of the annotations in the IAPR TC-12 collection is good and adequate for ImageCLEFphoto. | | | 3 | 3 | 1 |
| It was appropriate to provide a subset of the annotations (rather than the full set). | | | 4 | 1 | |
| The topics were realistic. | | 2 | 2 | 1 | 1 |

- Images with such good annotations are not at all real-life.

- I don't think that topics were realistic. I mean, if only titles are considered perhaps they are near to real search queries given by users but narratives are too long and tricky (several nested negations).
- I'm not sure what size of or what types of topics are ideal for the evaluation.
- As we are working in an environment of web based cross language image retrieval we found that query formulation was a little bit to "raffinate", sometimes with rather unusual terms.

# Feedback from survey (2006)
## *Selected results*

- Relevance Assessments & Performance Measures

| | strongly disagree | disagree | neutral | agree | strongly agree |
|---|---|---|---|---|---|
| The set of performance measure (MAP, P20, GMAP, BPREF) was adequate. | | | | 6 | 1 |
| MAP is a good indicator for of the effectiveness of an image retrieval system. | | 1 | 2 | 4 | |
| P20 should be adopted as the standard measure as many online image retrieval engines (Google, Yahoo, etc.) display by default 20 images on the first page of results. | | 2 | 3 | 1 | 1 |
| It is appropriate to use an interactive search and judge tool to complement the ground-truth with further relevant images that were not part of the pools. | | | 1 | 5 | 1 |

I don't think MAP to be a good indicator of the effectiveness of a retrieval system but I don't know other. The "feeling" of the user interacting with the system should be considered in some way but I don't know how (and I'm not sure if tests with 10 or 15 users, as in iCLEF experiments, are representative enough)

I don't think P20 is a good standard measure. Although most systems show 20 results in the first page I think that there are a lot of users that review more than 20. I don't have the proof of this assertion, of course, it is only a feeling.

I have no idea about this… but I prefer clear and simple performance measures which everybody knows and can optimise to. Suggestions: do it the standard way, that is: MAP

# Feedback from survey (2007)
## *Selected results*

- Image Annotations and Annotation Languages for 2007

| | strongly disagree | disagree | neutral | agree | strongly agree |
|---|---|---|---|---|---|
| English annotations. | | | 1 | 3 | 2 |
| German annotations. | | | 4 | 1 | 1 |
| Spanish annotations | | | 4 | 1 | 1 |
| Annotations with a randomly selected language (English, German, Spanish) because this is the most realistic scenario. | | 1 | 2 | 3 | |
| Visual features only, without annotations. | 1 | 2 | | 2 | 1 |

Other target languages: French

# Feedback from survey (2007)
## *Selected results*

- Topic Languages for 2007 topics

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| English | 5 | Portuguese | 1 | Norwegian | | Traditional Chinese | 1 |
| German | 2 | Dutch | | Finnish | | Monolingual only | |
| Spanish | 2 | Polish | 1 | Danish | | Images only | 1 |
| Italian | 2 | Russian | 1 | Japanese | 2 | | |
| French | 4 | Swedish | | Simplified Chinese | 1 | | |

Other query languages: Arabic

# Proposals for 2007

- Collection
  - Spanish annotations will be completed
  - Randomly select N images for each annotation language (and no annotation)
  - Assess “usefulness” of annotation fields
    - titles/keywords only vs. descriptions
- Resources
  - Runs from GIFT and VIPER for topic examples (other systems?)
  - Use 2006 data as training data

# Proposals for 2007

- Topics
  - "visual" topics part of the standard ad-hoc set
    - make visual properties part of the topic narrative?
  - Will provide 3 example images for each topic (but make sure these are removed from the database – not just the set of relevant images)
  - Support the languages suggested by participants
    - English, German, Spanish, Italian, French, Portuguese, Polish, Russian, Simplified/traditional Chinese, Arabic
  - Release topic narratives?
  - Creating topics – what/how?
    - realism vs. controlled parameters

# Proposals for 2007

- Tasks
  - ad-hoc retrieval task?
    - use same topics to see how much improvement can be gained one-year on
  - object recognition task?
  - submit at least one run using “off the shelf” tools?
- Performance measures
  - bridge gap between research and real-world
    - time taken for retrieval / cost of resources (and copyright!) / use of “off the shelf” resources
  - use measures which correlate with human satisfaction?
    - user satisfaction vs. system effectiveness
  - comparing systems
    - rank systems based on average rank from different measures


# Medical Ad-hoc Retrieval Task


# Automatic Annotation Tasks