

ImageCLEF / MUSCLE workshop

*Alicante 19/09/2006*

# ImagEVAL

## Usage-Oriented multimedia information retrieval Evaluation

Pierre-Alain Moëllic  
CEA List (France)

- Context of ImagEVAL
- ImagEVAL, what we try to do...
- Some details about the tasks
- Conclusion: ImagEVAL 2 ?

<http://www.imageval.org>

## Context of ImagEVAL

- Evaluation Campaigns
- French program TechnoVision

- Evaluation campaigns for information retrieval
  - Popularized by TREC
    - Text Retrieval Conference (first edition 92-93)
    - Year after year: diversification and multiplication of specific tracks
  - In French speaking countries... need to wait 1999 to have 2 TREC-like campaigns with some French databases (AMARYLLIS, CLEF)
  - Info. Retrieval evaluation has already been extended from purely textual retrieval to image / video retrieval :
    - TRECVID
    - ImageCLEF
  - Image retrieval without using text information (key words, captions) has been less explored. The «image retrieval» community need to make up for this delay !

- Standard TREC shared task test paradigm
  - Training corpus
  - Usually a test run
  - Test corpus given few months in advance
  - Requests given few weeks in advance
  - A fixed date to provide results
  - Evaluation of the first answers for each request of each participant to produce the pool of the expected best results
  - Recall/Precision curve and MAP on the 1000 first answers for each query

- Technological «vs» User-oriented
  - Technological evaluation
    - Establish a hierarchy between automatic processing technologies
    - The evaluation only considers the Recall / Precision result
  - User-oriented
    - Consider other end-users based characteristics for the evaluation
    - Some criteria :
      - Quality of the user interface
      - Response time
      - Indexing time
      - Adaptation time to a new domain
      - ...
    - Try to combine classical technological evaluation with end-user criteria
    - Make the end-users interact from the definition of the campaign, the creation of the ground truth to the final discussion and analysis

- Organizing such a campaign is a complex work needing appropriate resources and partnerships
- Some comments for possible changes...
  - ① **Training and test periods**
    - The more time and manpower you spend the best are your results...
    - Too important lag-time between the reception of data and posting the results usually implies extra testing and tuning that hardly represent the reality of a system
    - We should distinguish
      - Systems needing long training period
      - Systems can be tuned fastly
    - Idea from AMYRILLIS 2 : online and instantaneous participation.  
... but... only one participant !

### – ② Ground truths / Pooling technique

- Pooling technique: comparison of the pertinence of a document using a reference set composed of :
  - (1) Hand-verified documents = top rank doc returned by participants
  - (2) Unverified document
- If you find a unique good answer that is not in (1)... the document is considered as not relevant
- [Zobel, Sigir 98] : *“Systems that identify more new relevant documents that others get less benefit from the other contributors to the pool, and measurement to depth 1000 of these systems is likely to underestimate performance”*

### – ③ Size of the answer set

- Classical protocol : 1000 answer / query
- But a “real” end-user usually check the 20 first answers and rarely beyond... For a end-user, the “quality” of the beginning of the answer list is more important than the rest of the list



- TECHNO-VISION

- French program (ministries of Research & Defense)

<http://www.technologie.gouv.fr/technologie/infotel/technovision.htm>

- «... support the installation of a perennial infrastructure including the organization of evaluation campaigns and the creation of associated resources (data bases for the developments and the tests, metrics, protocols)»

- 10 evaluation projects

- 2 medical
- 2 video monitoring + 1 biometric (iris and face)
- 1 for technical and 1 for hand-written documents
- 2 for military applications
- 1 «generalist» : ImagEVAL

- 2 years to organize all the campaign: too short time !

- **28/02/2005** Steering Committee Meeting
- **T0+ 2**
- Metrics and protocols,
- Contracts with data providers,
- **29/03/2005** Consortium meeting.
- **05/2005**
- **Preparation of the learning and test run databases**
- **08/2005**
- **Sending of the learning databases**
- **Creation of the test run databases**
- **01/2006**
- **Test run evaluation**
- **Sending of the test run databases**
- **15/03/2006**
- **Participants: Sending of the results**
- **13/04/2006**
- **Results of the evaluations**
- **26/04/2006**
- **Consortium meeting. Discussion about the test run**
- **03/07/2006**
- **Official campaign**
- **Sending of the official databases**
- **21/08/2006**
- **Sending of the queries**
- **05/10/2006**
- **Participants : sendings the results**
- **03/11/2006**
- **Results of the evaluations**
- **End of 2006**
- **Contributions for the final conference**
- **11, 12, 13, 14 december 2006**
- **Workshop ImagEVAL**

- A consortium composed of 3 entities:
  - Steering Committee
    - Principal organizer : **NICEPHORE CITE**
    - Evaluation / organization: **TRIBVN**
    - Scientific animation : **CEA-LIST**
    - The steering committee:
      - Enables the construction and validation of the databases
      - Fixes the protocols (metrics,...)
      - Generates, analyses and diffuses the results
  - Data providers
  - Participants

- Data providers
  - Ensure the volume, the quality and the variety of the data
  - Privileged actors to discuss about the real needs
  - Data providers for ImagEVAL:
    - HACHETTE
    - RENAULT
    - National Museum Gathering (in french RMN)
    - CNRS (PRODIG) = Research Group for organization and diffusion of geographic information
    - Foreign Affair Ministry

- Some characteristic images



- We firstly had a lot of participants...
- Unfortunately, every TechnoVision projects met the problem of *"sorry we don't have manpower anymore..."*
- 2 explanations
  - Reality of the european research...
  - Participating to an evaluation in the computer vision community is CLEARLY NOT a priority nor a habit
- Finally we expect to keep 13 participants :
  - Labs
    - Mines de Paris (Fr)
    - INRIA – IMEDIA (Fr)
    - ENSEA – ETIS) (Fr)
    - University of Tours (RFAI) (Fr)
    - CEA-LIST – LIC2M) (Fr)
    - University of Strasbourg – LSIIT) (F)
    - University of Vienne – PRIP (Austria)
    - Hôpitaux universitaires de Genève (Switzerland)
    - University of Geneva – VIPER (Switzerland)
    - University of Barcelona (Spain)
  - Firms
    - Canon Research
    - LTU Tech
    - AdVestigo

## ImagEVAL

*What we try to do...*

- Main objectives of the first edition
- Choice of the tasks
- Constitution of the corpora
- Creation of the ground truth

- The main objectives of the first edition :
  - Constitute a pool of professional data provider and potential end-users
  - Participate to the emergence of an « evaluation culture » in the *image retrieval* and *image analysis* communities
  - Create a stable and robust technical base (metrics, protocols) for future tasks
  - Create and strengthen partnerships for future edition :  
TechnoVision program is not enough to organize a large scale and perennial evaluation



- Our first idea :
  - Organizing a big *Content Based Image Retrieval* evaluation
  - But it was not possible due to lack of time and manpower resources...
- Decide to break the complexity in several *shorter* tasks and asked professional and potential end-users what could be “interesting” tasks
  - Find objects or class of objects
  - Automatic classification or key-words generation
  - Protection of copyrights
  - Find pictures using a text/image mixed research

- For the 1<sup>st</sup> edition we try to follow some propositions hoping to follow all the propositions in future editions
  - Constitution of the databases
    - We aimed at building a diversified corpus covering the variety of usage of our commercial partners
    - Copyright problems were a real difficulty but agreements had been reached
    - It's one of the most important goal of ImagEVAL: establish a real cooperation between campaign organizers and data providers : important for the quality of the databases AND to spread the results to a large community
  - Ground truths
    - We decided to tag all the images of the databases
    - Two professionals (HACHETTE) realized the indexation. The ground truth creation has been made in a “end user” point of view. This point was also a strong decision of all the partners (second consortium meeting) that shows that the participants accept the idea of an end-user evaluation

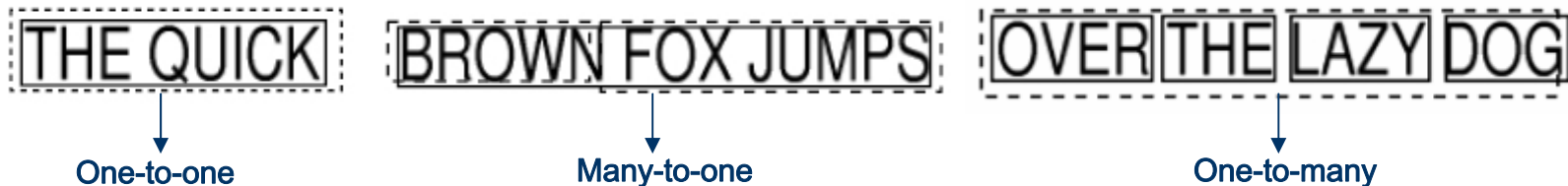
- Evaluation campaign
  - Because of the lack of experience of a lot of participants in evaluation campaigns we decided to organize a test run evaluation even if we don't have a lot of time
  - This test run was clearly profitable for everyone
  - Some participants were ready (and even asked) a very short time processing. That was very encouraging but it was not the unanimity so we decided – in order to keep enough participants !
    - to keep a standard delay (Queries / Results = 2 months)

## Some details about the tasks

- Metrics and protocol

### ● Metrics

- It's better to use well-known metrics even if it's not perfect than perpetually invent "the new best" metric...
- Except for task 3 that is more specific, we use Mean Average Precision and Recall / Precision analysis
- Mean Average Precision: 
$$\text{AveP} = \frac{\sum_{r=1}^N (P(r) \times \text{rel}(r))}{\text{number of relevant documents}}$$
- We use **TRECEVAL**
- Task 3 metric : Christian Wolf's metric, is based on Recall and Precision. A very intelligent metric that enables to treat on a same way different detection



<http://liris.cnrs.fr/christian.wolf/software/deteval/index.html>

- Invariance and robustness problems of image indexing technologies
- Important for copyright protection
- Test database
  - Kernel of N images. We applied 16 transformations (could be combined)
    - Geometric transformations (rotation, projection...)
    - Chromatic transformations (saturation, b&w, negative, ...)
    - Structural transformations (border, text adding...)
    - Others... (JPEG quality, blur, noise,...)
  - Test run : about 4500 images (N=250)
  - Official test : about 45000 images (N=2500)
- 2 sub-tasks
  - 1.1
    - From a kernel image, retrieve all the transformed images
    - 50 queries. 50 answer / query
  - 1.2
    - From a transformed image, find the kernel image
    - 60 queries. 50 answers / query




- **Metrics**

- **1.1**

- **MEAN AVERAGE PRECISION (MAP)**
    - Recall / Precision
    - TRECEVAL

- **1.2**

- Mean Reciprocal Rank (MRR)
    - Provided by CEA List

- Image retrieval for Internet application
- Database : web pages in French
  - Text / Image Segmentation with a tool proposed by CEA
  - Run test : 400 URL
  - Official test : 700 URL
  - The database was composed using common “encyclopaedic” queries :
    - Geographic site
    - Objects
    - Animals, ...
  - We also use Wikipedia
- Objective
  - Retrieve all the images answering a query : <request> + example images
  - Example: <Statue de la Liberté> +
- Queries
  - Test run :
    - 15 queries
    - 150 answers / query
  - Official test :
    - 25 queries
    - 300 answers / query



- **Data**

- Example of a text file (using the segmentation tool...)

[17][187\_ayers-rock.jpg]

### Introduction

Le Parc National d'Uluru est l'un des sites touristiques les plus visités, quasi centre géographique de l'Australie dans le Sud du Territoire du Nord. Il est le siège du célèbre Ayers Rock , lieu sacré et symbole des légendes aborigènes. L'Ayers Rock est le plus gros rocher à la surface de la terre, avec sa circonférence de 9.5 km et sa hauteur de 348 m. Tel un iceberg, seul un dixième de sa masse émerge. Il est situé à 400 km au sud d'Alice Springs et 2000 km au sud de Darwin. Ayers Rock est un site hors du temps, il a environ 600 millions d'années!

Connu des aborigènes depuis toujours, ce n'est qu'en 1872 qu'un européen l'a aperçut pour la première fois. Le site porte le nom du Premier ministre d'Australie Méridionale, Henry Ayers , nom donné sans savoir que le rocher était déjà connu des Aborigènes sous le nom d'Uluru (prononcez "oulourrou" en roulant le "r").

Ce monument est inscrit au Patrimoine mondial de l'Unesco depuis 1987. Les tribus aborigènes Anangus à qui appartient la région d'Ayers Rock en ont reçu la propriété officielle et définitive en octobre 1985. Ils l'ont louée en retour au gouvernement pour 99 ans.

Selon la direction de la lumière, ce monolithe prend différentes couleurs telles que rouge ou marron, il vire au carmin puis au violet au coucher du soleil, revêtant un aspect si poétique et envoûtant. Mais Uluru est avant tout un lieu sacré pour les aborigènes. Ils lisent des histoires dans son paysage et vous les racontent si vous prenez le temps d'écouter.

### Géographie

Lorsque l'Australie se désolidarise du continent préhistorique Gondwana, le phénomène se produit avec un minimum de secousses sismiques et autres drames telluriques, ce qui fait que le continent est grosso modo inchangé depuis six cent millions d'années.

La terre est fortement hostile : 43% de la surface se compose de désert ou de terres arides, 20 % de terres semi-arides, et 7% de roche nue. Plus des deux tiers du continent ne sont pas favorable à la vie.

[18][[grotte2.jpg] Cliquez pour agrandir-[19][187\_small\_grotte2.jpg] Les sous-sol sont riches en mines d'or, de pierres précieuses, de minerais et d'uranium (n°1 mondial pour le fer et la bauxite).

Localisation of the pictures with the name



- **Metric**

- **MEAN AVERAGE PRECISION (MAP)**
- **Recall / Precision**

- Even if it's a very experimental task, it was clearly the most difficult task to organize
- The test is interesting but we will need for ImagEVAL 2 to build a more robust database
  - Not only French web sites
  - Use XML structure for text information
  - Use other data :
    - Press article

- **Task 3. Text detection in an image**
- **Database**
  - Old post cards with captions
  - Indoor and outdoor pictures with text as scene elements
- **Objective**
  - Detect and localize text areas in all the images of the database
- **Queries**
  - Test run:
    - 500 images
  - Official test:
    - 500 images



## Task 3 Text detection in an image

- Text area is characterized by a bounding box [(X1,Y1) (X2,Y2)]

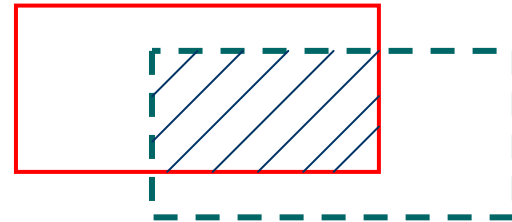
- Metric**

- ICDAR based

- Based on recall and precision

- $R = \text{Aire\_inter} / \text{Aire\_vt}$

- $P = \text{Aire\_inter} / \text{Aire\_res}$



- Metric developed C.Wolf (INSA Lyon)

- Amelioration of the ICDAR metric. This metric enables a better evaluation of the bounding boxes merging problems
    - Christian Wolf enables to deal with one-to-one / many-to-one / one-to-many matching

THE QUICK

One-to-one

BROWN FOX JUMPS

Many-to-one

OVER THE LAZY DOG

One-to-many

- **Task 4 : object detection**

- **Database**

- 10 objects or class of objects

*Tree*

*Minaret*

*Eiffel Tower*

*Cow*

*American flag*

*Car*

*Armored vehicle*

*Sun glasses*

*Road signs*

*Plane*

- Learning database / Dictionary database about 750 images
- Test run : 3 000 images
- Official test : 15 000 images

- **Objective**

- Find all the image containing the request object
- Example of a query : <Eiffel Tower>

- **Run**

- The first run only uses the learning data
- Supplementary data could be used for other runs. Nature and volume will be described

- **Queries**

- Run test
  - 4 objects
  - 500 answer / request
- Official test
  - 10 objects
  - 5000 answer / request

## Task 4 Objects detection

- Examples of images



Dictionnary database



Test database

- Metrics:

- MEAN AVERAGE PRECISION (MAP)
- Recall / Precision

- **Task 5. Semantics extraction**
- **Database**
  - About 10 attributes :  
*B&W pictures, Color pictures, Colorized B&W, Art reproduction, Indoor, Outdoor, Day, Night, Nature, Urban*
  - Learning database 5000 images
  - Run test : 3 000 images
  - Official test : 30 000 images
- **Objective**
  - Find all the image corresponding to an attribute or a series of attributes
  - Example of a request : Color / Outdoor / Day / Urban
- **Run**
  - The first run only uses the learning data
  - Supplementary data could be used for other runs. Nature and volume will be described
- **Requests**
  - Test run :
    - 5 attributes or lists of attributes
    - 1000 answers / request
  - Official test :
    - 13 attributes or list of attributes
    - 1000 answers / request



## Task 5 Semantic extraction

(1) Color	1
(2) Black White	0
(3) Colorized Black White	0
(4) Art reproduction	0
(5) Indoor	0
(6) Outdoor	1
(7) Night	0
(8) Day	1
(9) Natural	0
(10) Urban	1



(C) Editing

- **Metrics**

- **MEAN AVERAGE PRECISION (MAP)**
- Recall / Precision



## Conclusion

### ImagEVAL 2 ?

- Is an ImagEVAL 2 possible ?
- What we learn...
- Some changes for the second edition

- We don't have any idea if TechnoVision will continue...
- CEA List wants to continue ImagEVAL :
  - Open the campaign to (more) European participants
  - Change and enlarge the Steering Committee to ameliorate the organization
  - Propose a more complete website that should enable :
    - A platform to download large databases
    - A live platform evaluation : the participant directly upload the answer file and receive the results
  - Organize new tasks
    - The task 2 (mixed text/image research) is not enough, we need to imagine a bigger, more robust and realistic database

- Too early to draw lessons from ImagEVAL but...
  - The scientific community is receptive
  - Involvement of important data provider and potential end users (HACHETTE, Renault, Museums...) is clearly encouraging
  - We learned a lot about the organization of a campaign and – above all – we manage to get in touch with a lot of people that are ready to continue our efforts