MediaEval Benchmark 2013

MediaEval Benchmarking Initiative for Multimedia Evaluation

The "multi" in multimedia: speech, audio, visual content, tags, users, context

Retrieving Diverse Social Images Task

- task overview -

Bogdan Ionescu (UPB, Romania)
María Menéndez (UNITN, Italy)
Henning Müller (HES-SO in Sierre, Switzerland)
Adrian Popescu (CEA LIST, France)

University Politehnica of Bucharest

UNIVERSITY OF TRENTO - Italy

October 18-19, Barcelona, Spain

Outline

- The Retrieving Diverse Social Images Task
- Dataset and Evaluation
- Participants
- Results
- Discussion and Perspectives

Diversity Task: Motivation

Objective: the task addresses the problem of **result diversification** in the context of *social photo retrieval*.

Use case: we consider a tourist use case where a person tries to find more information about a place she is potentially visiting. The person has only a vague idea about the location, knowing the name of the place.

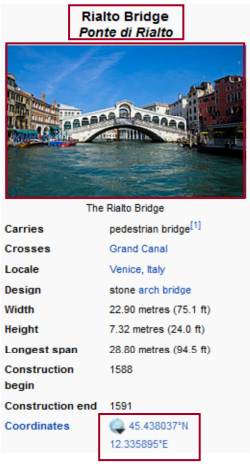
… e.g., looking for **Rialto Bridge** in Italy

Diversity Task: Motivation

Objective: the task addresses the problem of **result diversification** in the context of *social photo retrieval*.

Rialto Bridge

-> get more information from Wikipedia,



Diversity Task: Motivation

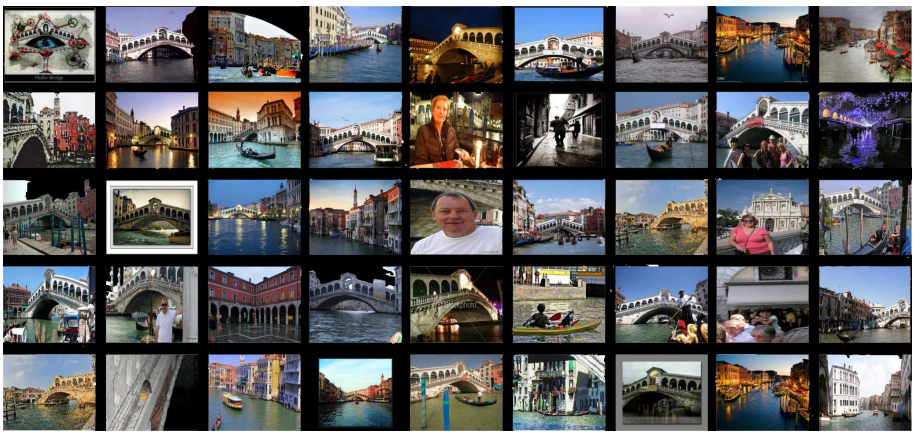
… now, how to get some more accurate photos ?

Google Panoramio Picasa flickr

… query using text and GPS tags:
“Rialto Bridge”,
45.438037ºN, 12.335895ºE

browse the results …

Diversity Task: Motivation



page 1

Diversity Task: Motivation

page n

Diversity Task: Motivation

… too many results to process,

inaccurate, e.g., people in focus, other views or places

meaningless objects

redundant results, e.g., duplicates, similar views …

Diversity Task: Motivation

page 1

Diversity Task: Motivation

page n

Diversity Task: Definition

Participants receive a *ranked list* of photos with locations retrieved from Flickr using its default "relevance" algorithm.

Goal of the task: *refine* the results by providing a ***ranked*** list of *up to 50 photos* (***summary***) that are considered to be both ***relevant*** and ***diverse*** representations of the query.

relevant*: common visual representation of the location, e.g., different views at different times of the day/year, inside views, close-ups, drawings, sketches, creative views, which contains partially or entirely the location.

diverse*: depicting different visual characteristics of the location, with a certain degree of complementarity, i.e., most of the perceived visual information is different from one photo to another.

*we thank the participants to the task survey for their precious feedback on these definitions.

Diversity Task: Target

going from this …

… to something like this

Diversity Task: Definition

The task builds on current technology rather than requesting participants to develop their own retrieval systems
e.g., [ImageCLEF Photo Retrieval 2009]

Participants are submitting up to 5 runs:

- **required runs**:
  run 1: automated using *visual information only;*
  run 2: automated using *textual information only*;
  run 3: automated using *textual-visual* fused without other resources than provided by the organizers;
- **general runs**:
  run 4: *human-based* or *hybrid human-machine* approaches;
  run 5: *everything allowed* including using data from external sources (e.g., Internet).

Dataset: Statistics

The dataset consists of **396 landmark locations** (natural or man-made, e.g., sites, museums, monuments, buildings, roads, bridges) unevenly spread over 39 countries around the world:

Dataset: Statistics

Each location contains:

- the location name & GPS coordinates;
- a link to its Wikipedia web page;
- a representative photo from Wikipedia;
- a ranked set of Creative Commons photos retrieved from Flickr (up to 150 photo/location);
- metadata from Flickr (e.g., tags, description, views, #comments, date-time photo was taken, user, etc);
- some general purpose visual and text content descriptors.

Retrieval method (we use Flickr API):

- using the location name as query (**keywords**);
- using location name and GPS coordinates* (**keywordsGPS**).

* we use a 1 Km radius around the GPS coordinates.

Dataset: Statistics

Basic statistics:

- **devset** (intended for designing and validating the methods)

	devset		
	#locations	*#images*	*min-avg.-max img./location*
keywords	25	2,281	30 - 91.2 - 150
keywordsGPS	25	2,837	45 - 113.5 - 150
overall	50	5,118	30 - 102.4 - 150

- **testset** (intended for final benchmark)

	testset		
	#locations	*#images*	*min-avg.-max img./location*
keywords	135	13,591	30 - 100.7 - 150
keywordsGPS	211	24,709	35 - 117.1 - 150
overall	346	38,300	30 - 110.7 - 150

$\Rightarrow$ total number of images: 43,418.

Dataset: Ground Truth

Relevance and diversity annotation was carried out by:

- **expert annotators***
  - *devset*: relevance (6 annotations), diversity (1 annotation issued from 3 experts);
  - *testset*: relevance (3 annotations issued from 7 expert annotators), diversity (1 annotation from 4 expert annotators);
  - lenient majority voting.
- **crowd workers****
  - relevance (3 annotations) and same majority voting;
  - diversity (3 annotations).

* have advanced knowledge of the location characteristics.

** crowd annotation was performed for a selection of 50 testset locations via CrowdFlower.

Dataset: Ground Truth

Basic annotation statistics:

- **expert annotations**

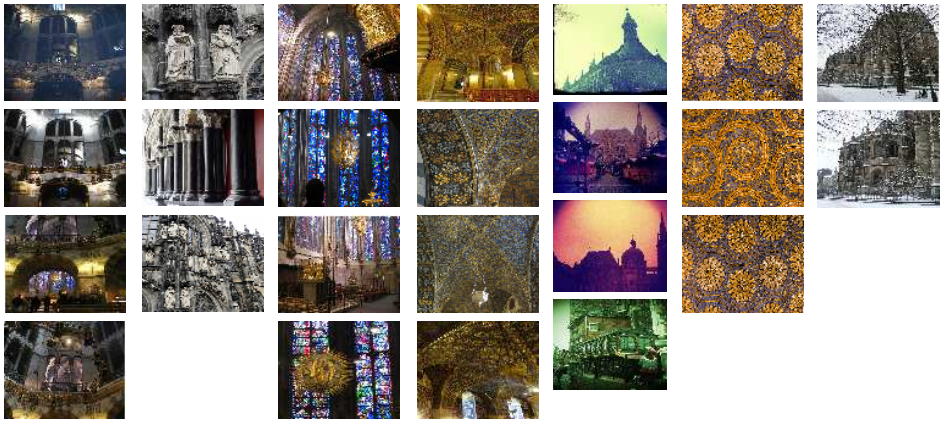
devset		**testset**	
keywords	*keywordsGPS*	*keywords*	*keywordsGPS*
relevance (avg. Kappa / % relevant img.)			
0.68 68%	0.61 79%	0.86 55%	0.75 75%
diversity (avg. clusters per location / avg. img. per cluster)			
10.4 5.5	12.8 7.4	11.8 4.2	14.5 5.8

- **crowd annotations**

testset (selection of 50 locations, 6169 photos)		
relevance (avg. Kappa and % relevant img.): 0.36 69%		
diversity (avg. clusters per location / avg. img. per cluster)		
GT1	*GT2*	*GT3*
3.5 43.1	4.3 30.4	6.3 24

Dataset: Ground Truth

Diversity expert annotation example (Aachen Cathedral*, Germany):



chandelier | architectural details | stained glass windows | archway mosaic | creative views | close up mosaic | outside winter view

* excerpt, the total number of clusters is 15.

Evaluation

Official metrics:

official ranking CR@10

- **Cluster Recall* @ X = Nc/N** (CR@X)

where X is the number of ranked images, N is the total number of clusters for the current location (from ground truth, N<=20) and Nc is the number of different clusters represented in the X ranked images;

- **Precision @ X = R/X** (P@X)

where R is the number of relevant images;

- **F1-measure @ X = harmonic mean of CR and P** (F1@X)

Metrics are reported for different values of X (5,10,20,30,40 and 50) on per location basis as well as overall (average).

* cluster recall is computed only on the relevant images.

** official metrics were computed on testset by excluding locations (ids) 81, 298, 305 and 367 for which there were no relevant images in the ground truth.

Participants: Basic Statistics

- **Survey (February 2013):**
  - 55 respondents were interested in the task (23 very interested);
- **Registration (May 2013):**
  - 24 teams registered from 18 different countries (3 teams are organizer related);
- **Crossing the finish line (September 2013):**
  - 11 teams finished the task (8 countries) including 3 organizer related teams and 1 late submission;
  - 38 runs were submitted from which 2 brave human-machine!
- **Workshop participation (October 2013):**
  - 8 teams are represented at the workshop.

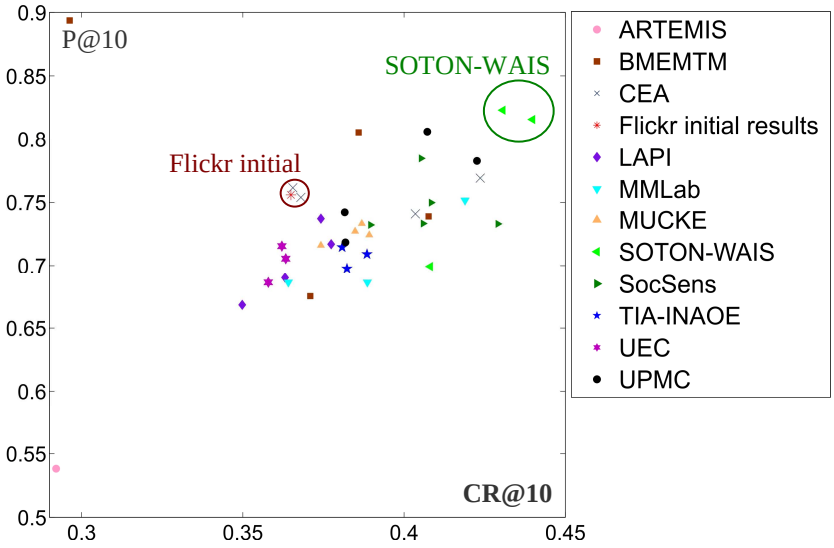
Participants: Approaches

team	*country*	*1-visual*	*2-text*	*3-text-visual*	*4-human*	*5-free*
SocSens	UK	√	√	√	hybrid	Exif, weather
SOTON-WAIS	UK	√	√	√	x	x
MUCKE*	Turkey	√	√	√	x	text-visual
LAPI*	Romania	√	√	√	x	visual
TIA-INAOE	Mexico	√	√	√	x	x
UEC	Japan	√	√	√	x	x
BMEMTM	Hungary	√	√	√	human only	
UPMC	France	√	√	√	x	visual-text
ARTEMIS**	France	√	x	x	x	x
CEA*	France	√	√	√	x	user date
MMLab	Belgium	√	√	√	x	x

* organizer related team.

** late submission ☺.

Results: expert ground truth



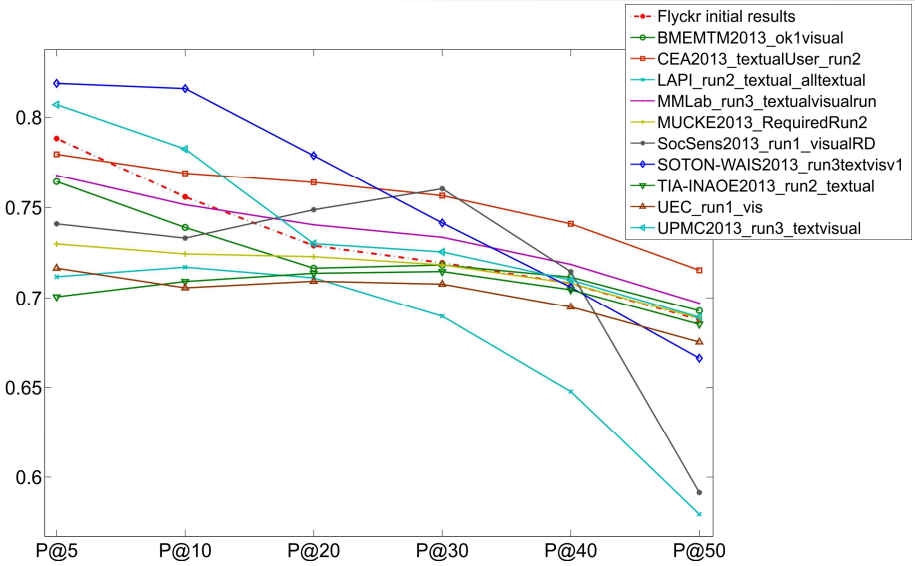
* all participant runs, evaluation on the entire dataset (keywords + keywordsGPS).

Results: expert ground truth

team/run	**P@10**	P@20	**CR@10**	CR@20	**F1@10**	F1@20
SOTON-WAIS2013_run3textvisv1	**0.8158**	0.7788	**0.4398**	0.6197	**0.5455**	0.6607
SocSens2013_run1_visualRD	**0.733**	0.7487	**0.4291**	0.6314	**0.5209**	0.6595
CEA2013_textualUser_run2	**0.769**	0.7639	**0.4236**	0.6249	**0.5227**	0.6593
UPMC2013_run3_textvisual	**0.7825**	0.73	**0.4226**	0.6268	**0.53**	0.6498
MMLab_run3_textualvisualrun	**0.7515**	0.7404	**0.4189**	0.6236	**0.5174**	0.6514
BMEMTM2013_ok1visual	**0.7389**	0.7164	**0.4076**	0.6139	**0.5066**	0.6363
MUCKE2013_RequiredRun2	**0.7243**	0.7228	**0.3892**	0.5749	**0.4905**	0.6182
TIA-INAOE2013_run2_textual	**0.7091**	0.7136	**0.3885**	0.5732	**0.4801**	0.6102
LAPI_run2_textual_alltextual	**0.717**	0.7111	**0.3774**	0.5734	**0.4736**	0.6078
Flyckr initial results	**0.7558**	0.7289	**0.3649**	0.5346	**0.4693**	0.5889
UEC_run1_vis	**0.7056**	0.7092	**0.3633**	0.5448	**0.4617**	0.5926
ARTEMIS2013_av1_reloaded5	**0.5383**	0.3379	**0.2921**	0.3306	**0.3653**	0.3194

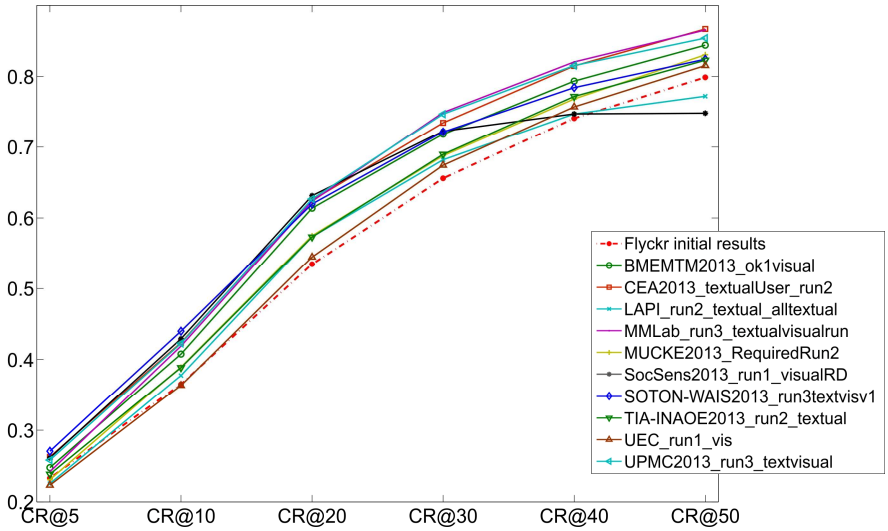
* team best runs according to CR@10, assessed on the entire dataset (keywords + keywordsGPS).

Results: expert ground truth



* team best runs according to CR@10, assessed on the entire dataset (keywords + keywordsGPS).

Results: expert ground truth



* team best runs according to CR@10, assessed on the entire dataset (keywords + keywordsGPS).

Results: crowd ground truth

team/run	**P@10**	P@20	**CR@10**	CR@20	**F1@10**	F1@20
UPMC2013_run3_textvisual	**0.7490**	0.6867	**0.7880**	0.8745	**0.7421**	0.7495
MMLab_run1_visualrun	**0.7245**	0.7061	**0.7721**	0.8789	**0.7155**	0.7603
SocSens2013_run1_visualRD	**0.7286**	0.7653	**0.7636**	0.8865	**0.7235**	0.8020
LAPI_run3_textual_visual_prob&CSD	**0.6796**	0.6929	**0.7515**	0.8653	**0.6675**	0.7440
MUCKE2013_RequiredRun3	**0.7245**	0.7102	**0.7503**	0.8644	**0.7050**	0.7559
CEA2013_multimedia_run3	**0.7673**	0.7724	**0.7484**	0.8354	**0.7268**	0.7768
TIA-INAOE2013_run3_multimedia	**0.6714**	0.6918	**0.7480**	0.8675	**0.6769**	0.7464
SOTON-WAIS2013_run1visonlyv1	**0.6612**	0.6827	**0.7477**	0.8803	**0.6707**	0.7482
BMEMTM2013_ok3textvis	**0.6469**	0.6551	**0.7477**	0.8616	**0.6597**	0.7206
UEC_run2_text	**0.6673**	0.6847	**0.7331**	0.8429	**0.6659**	0.7366
Flyckr initial results	**0.6816**	0.7061	**0.6643**	0.8119	**0.6269**	0.7186
ARTEMIS2013_av1_reloaded5	**0.6449**	0.4112	**0.7510**	0.7872	**0.6615**	0.5128

* team best runs according to CR@10, assessed on all the three crowd ground truth (average).

Results: human ranking

3 persons were asked to rank all the run results according to their own judgment of visual relevance and diversity.

Asinelli Tower, Italy
(in general high diversity but variable relevance)

team/run	average score
SOTON-WAIS2013_run2textonlyv2	**1.67**
LAPI_run1_visual_HOG	**4.00**
SocSens2013_run1_visualRD	**4.00**
UEC_run3_mix	**8.67**
UPMC2013_run2_text	**8.67**
BMEMTM2013_ok3textvis	**9.33**
MMLab_run3_textualvisualrun	**9.33**
TIA-INAOE2013_run2_textual	**12.33**
MUCKE2013_RequiredRun5	**14.33**
CEA2013_multimedia_run3	**18.00**

Arc de Triomf, Spain
(in general high relevance but low diversity)

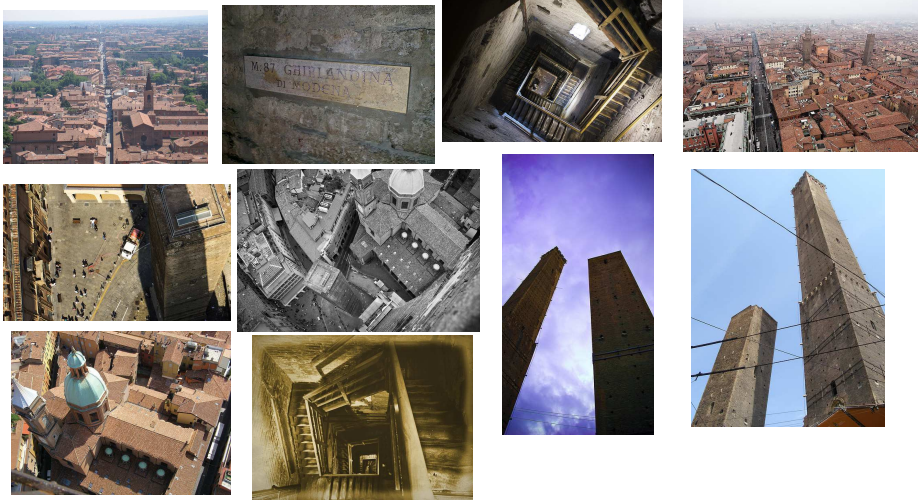
team/run	average score
SocSens2013_run1_visualRD	**1.33**
TIA-INAOE2013_run2_textual	**3.67**
SOTON-WAIS2013_run2textonlyv2	**5.33**
CEA2013_textualUserDate_run5	**5.67**
LAPI_run1_visual_HOG	**7.67**
MMLab_run1_visualrun	**10.00**
MUCKE2013_RequiredRun5	**10.67**
UPMC2013_run1_visual	**12.33**
BMEMTM2013_ok1visual	**13.67**
UEC_run1_vis	**23.00**

* team best runs according to the average visual score.

Results: human ranking

Asinelli tower

Flickr initial results

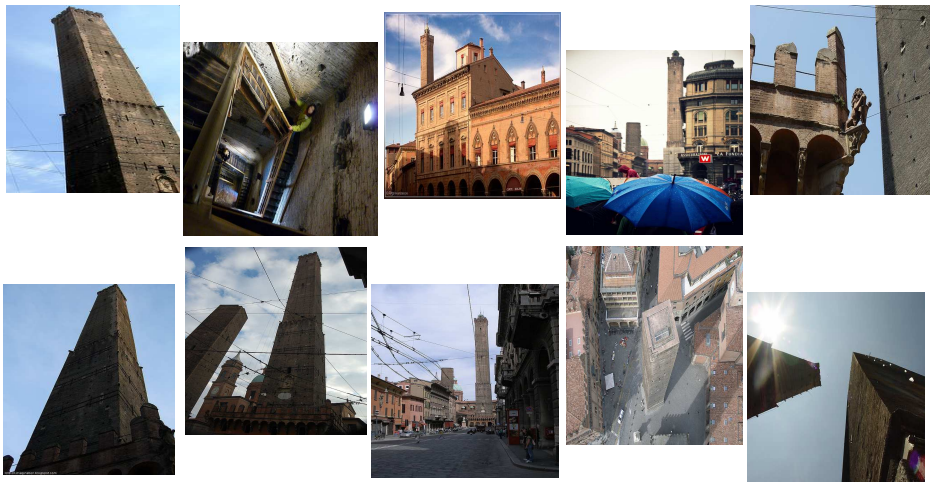


29

Results: human ranking

Asinelli tower

SOTON-WAIS2013_run2textonlyv2 (highest rank 1.9)



30

Results: human ranking

Arc de Triomf

Flickr initial results

Results: human ranking

Arc de Triomf

SocSens2013_run1_visualRD (highest rank 1.33)

Discussion

Methods:

- graph representations, re-ranking, optimization approaches, data clustering, human-based or hybrid (machine-human);
- best run @10: re-ranking + Greedy Min-Max similarity diversifier & using both visual and text information (SOTON-WAIS);
- not a big overall improvement (~10%), results are close to the actual technology - we should aim for high CR (>90%).

Dataset:

- mining for Creative Commons increases artificially the diversity;
- keywordsGPS is more accurate than keywords alone;
- evaluation depends on the ground truth (however, regardless the ground truth there is a similar improvement over the baseline);
- descriptors proved to be very useful.

Present & Perspectives

For 2013:

- we believe that the task (Brave Task) was a success!
- testset ground truth is to be released to participants (soon);
- the entire dataset is to be made publicly available (soon).

For 2014:

- simplify the task thus to have only 2-3 clusters per location and thus to facilitate achieving high cluster recall;
- a different use case, more images per location, …

Acknowledgements: many thanks to the task supporters for their precious help: Anca-Livia Radu, Bogdan Boteanu, Ivan Eggel, Sajan Raj Ojha, Oana Pleș, Ionuț Mironică, Ionuț Dută, Andrei Purică, Macovei Corina and Irina Nicolae.

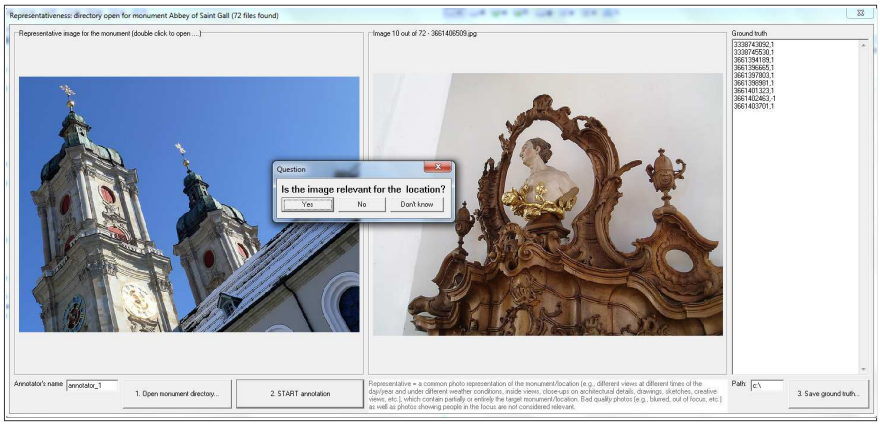
Questions & Answers

Thank you!

… and please contribute to the task by uploading free Creative Commons photos on social networks! ☺

Dataset: Ground Truth

Relevance and diversity annotation was carried out by **experts** as well as by **crowd** workers*.

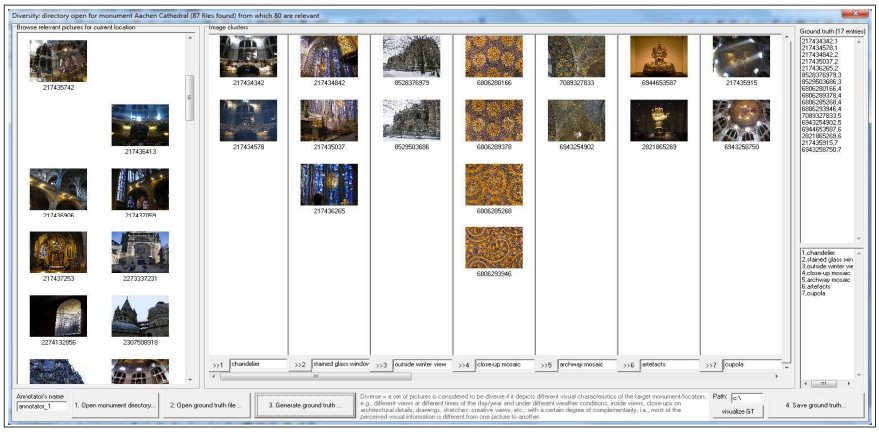


relevance tool

* crowd annotation was performed for a selection of 50 locations on CrowdFlower.

Dataset: Ground Truth

Relevance and diversity annotation was carried out by **experts** as well as by **crowd** workers*.



diversity tool

* crowd annotation was performed for a selection of 50 locations.