MediaEval Benchmark 2014

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)
Adrian Popescu (CEA LIST, France)
Mihai Lupu (TUW, Austria)
Henning Müller (HES-SO in Sierre, Switzerland)

University Politehnica of Bucharest

cea **List**

TU WIEN TECHNISCHE UNIVERSITÄT WIEN Vienna University of Technology

Hes·so VALAIS WALLIS

October 16-17, Barcelona, Spain

Outline

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

Diversity Task: Objective & Motivation

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

Why diversifying search results?
- a method of tackling queries with unclear information needs;
- queries involve many declinations, e.g., sub-topics;
- widens the pool of possible results and increases the system performance;
…

**Relevance and Diversity** (~antinomic):
too much *diversification* may result in losing relevant items while increasing only the *relevance* will tend to provide near duplicate information.

3

Diversity Task: Objective & Motivation #2

The concept appeared initially for text retrieval but regains its popularity in the context of multimedia retrieval.
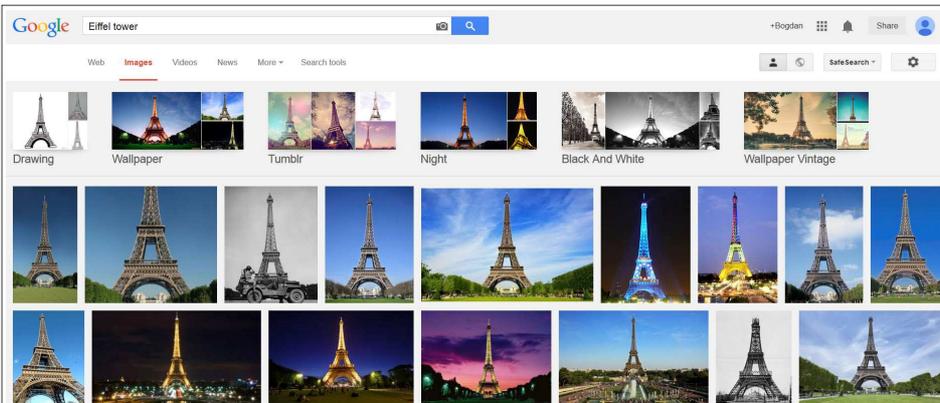
[Google Image Search for "Eiffel tower", 12-10-2014]

4

Diversity Task: Use Case

To disambiguate the diversification need, we introduced a very focused use case scenario …

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: Use Case #2

… learn more information from Wikipedia

Diversity Task: Use Case #3

… how to get some more accurate photos ?

query using text "Rialto Bridge" …

… browse the results

Diversity Task: Use Case #4

page 1

Diversity Task: Use Case #5

page n

Diversity Task: Use Case #6

… 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: Use Case #7

page 1

Diversity Task: Use Case #8

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 photo representation of the location, e.g., different views at different times of the day/year and under different weather conditions, inside views, close-ups, drawings, sketches, creative views, which contain partially or entirely the target 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 task survey respondents for their precious feedback on these definitions.

13

Diversity Task: Target

going from this …

…
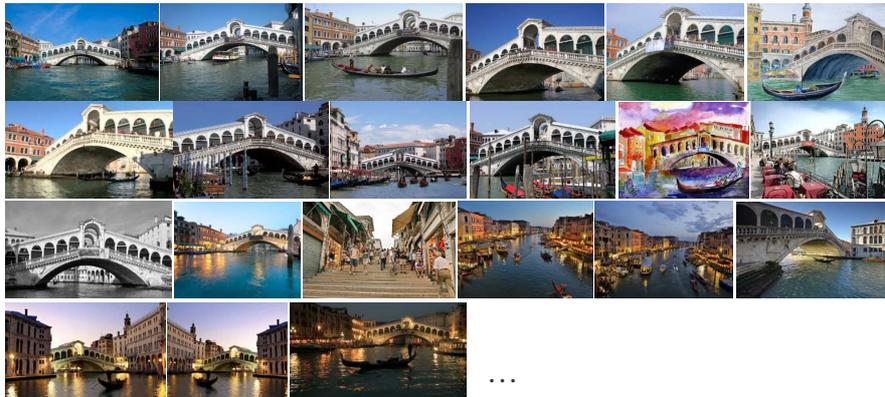
14

Diversity Task: Target

… to something like this:

Dataset: General Information

The dataset consists of **300 landmark locations** (natural or man-made, e.g., sites, museums, monuments, buildings, roads, bridges) unevenly spread over 35 countries around the world:

[Google Maps ©2014 MapLink]

Dataset: Resources

Location information consists of:
- the location name & GPS coordinates;
- a link to its Wikipedia web page;
- up to **5 representative photos** from Wikipedia;
- a ranked set of Creative Commons photos retrieved from Flickr (up to **300 photos per location**);
- metadata from Flickr (e.g., tags, description, views, #comments, date-time photo was taken, username, **userid**, etc);
- some general purpose visual and text content descriptors;
- **an automatic prediction of user annotation credibility**;
- relevance and diversity ground truth (**up to 25 classes**).

Retrieval method (we use Flickr API):
- use of the location name as query.

[2014: more focus on social aspects]

* the differences compared to 2013 data are depicted in bold.

Dataset: User Credibility

Idea: give an automatic estimation of the quality of tag-image content relationships;

~ indication about which users are most likely to share relevant images in Flickr (according to the underlying task scenario).

- **visualScore**: for each Flickr tag which is identical to an ImageNet concept, a classification score is predicted and the visualScore of a user is obtained by averaging individual tag scores;

- **faceProportion**: the percentage of images with faces out of the total of images tested for each user;

- **uploadFrequency**: average time between two consecutive uploads in Flickr;

…

Dataset: Statistics

Some basic statistics:

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

#locations	#images	min-average-max img. per location
30	8,923	285 - 297 - 300

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

#locations	#images	min-average-max img. per location
123	36,452	277 - 296 - 300

$\Rightarrow$ total number of provided images: 45,375.

- **credibilityset** (intended for training/designing credibility desc.)

#locations	#images*	#users	average img. per user
300	3,651,303	685	5,330

* images are provided via Flickr URLs.

Dataset: Ground Truth

Relevance and diversity annotations were carried out by **expert annotators***:

- **devset**: relevance (3 annotations), diversity (1 annotation issued from 2 experts + 1 final master revision);

- **testset**: relevance (3 annotations issued from 11 expert annotators), diversity (1 annotation from 3 expert annotators + 1 final master revision);

- **credibilityset**: only relevance for 50,157 photos (3 annotations issued from 9 experts);

- lenient majority voting for relevance.

* advanced knowledge of location characteristics mainly learned from Internet sources.

Dataset: Ground Truth #2

Some basic statistics:

- **devset**:

relevance	Kappa agreement* **0.85**	% relevant img. 70
diversity	avg. clusters per location 23	avg. img. per cluster 8.9

- **testset**:

relevance	Kappa agreement* **0.75**	% relevant img. 67
diversity	avg. clusters per location 23	avg. img. per cluster 8.8

- **credibilityset**:

relevance	Kappa agreement* **0.75**	% relevant img. 69

*Kappa values > 0.6 are considered adequate and > 0.8 are considered almost perfect.

Dataset: Ground Truth #3

Diversity annotation example (Aachen Cathedral, Germany):

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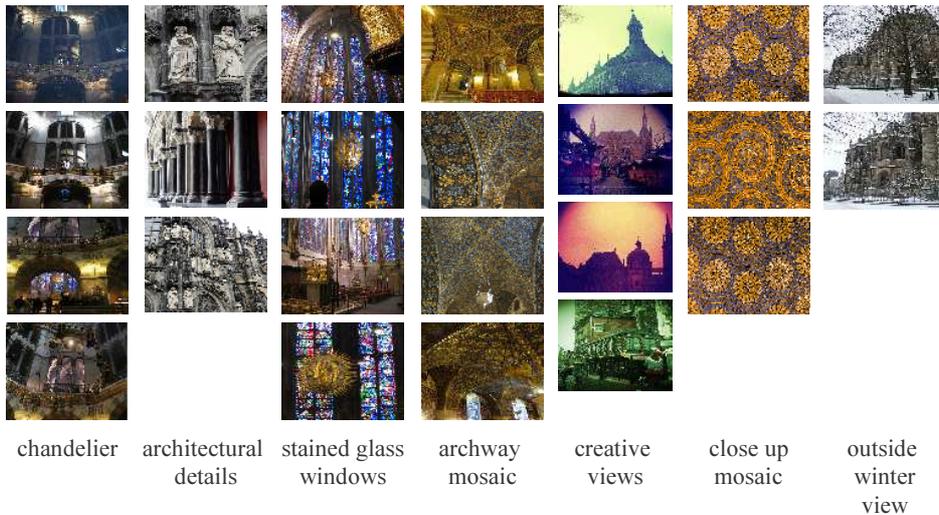
Evaluation: Required Runs

Participants are required to submit 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: automated using *credibility information*;
  run 5: *everything allowed,* e.g., human-based or hybrid human-machine approaches, including using data from external sources (e.g., Internet).

Evaluation: Official Metrics   official ranking F1@20

- **Cluster Recall*** **@ X = Nc/N** (CR@X)
  where X is the cutoff point, N is the total number of clusters for the current location (from ground truth, $N \leq 25$) 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 for the relevant images.

Participants: Basic Statistics

- **Survey (February 2014):**
  - 66 (55) respondents were interested in the task, **26** (23) very interested;

- **Registration (April 2014):**
  - 20 (24) teams registered from **15** (18) different countries (3 teams are organizer related);

- **Crossing the finish line (September 2014):**
  - 14 (11) teams finished the task, **12** (8) countries, including 3 organizer related teams (no late submissions);
  - 54 (38) runs were submitted from which **1** (2) brave human-machine!

- **Workshop participation (October 2013):**
  - 10 (8) teams are represented at the workshop.

* the numbers in the brackets are from 2013.

Participants: Submitted Runs

team	country	1-visual	2-text	3-text-visual	4-cred.	5-free
BU-Retina	Turkey	√	x	x	x	visual
CEALIST*	France, Austria	√	√	√	√	visual+cred.
DCLab	Hungary	√	√	√	√	multimodal
DWS	Germany	√	√	x	x	x
LAPI*	Romania, Italy	√	√	√	√	human-mach.
miracl	Tunisia	√	x	x	x	visual
TUW*	Austria	√	√	√	√	multimodal
MIS	Austria	√	√	√	x	visual
MMLab	Belgium, S. Korea	√	√	√	x	visual-text
PeRCeiVe@UNICT	Italy	√	√	√	x	visual
PRa-MM	Italy	√	√	√	√	multimodal
Recod	Brazil	√	√	√	√	multimodal
SocSens	Greece	√	√	√	x	visual-text
UNED	Spain	x	√	x	x	text

* organizer related team.

Results: P vs. CR @20 (all runs)
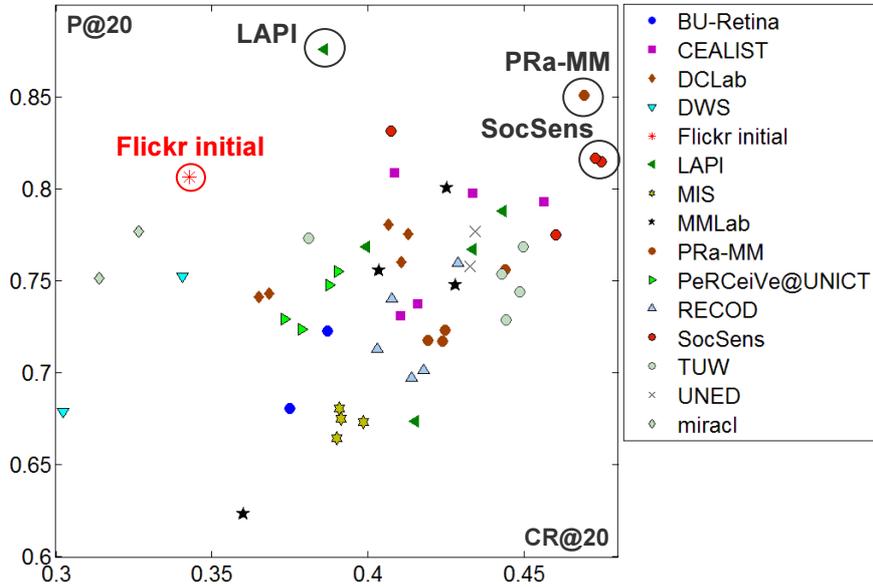
Results: Official Ranking According to F1@20

> team best runs only (full ranking will be sent via email);

team/run	P@10	P@20	CR@10	CR@20	F1@10	F1@20
PRa-MM2014_run5_Final	**0.8659**	**0.8512**	0.2976	0.4692	0.4362	**0.5971**
SocSens2014_run5	0.8268	0.815	**0.3027**	**0.4747**	**0.4394**	0.5943
CEALIST2014_run_5_general	0.7951	0.7931	0.2803	0.4563	0.4076	0.571
TUW2014_RUN1_visual_new.test	0.7984	0.7687	0.2827	0.4497	0.4124	0.5602
LAPI2014_run2_HC-RF_text_TF	0.7984	0.7882	0.2661	0.4431	0.3928	0.5583
Run5_UNED2014all_results	0.7748	0.7772	0.2679	0.4343	0.3932	0.5502
MMLab2014_run5_useall	0.7967	0.8008	0.2508	0.4252	0.3748	0.5455
RECOD2014_4vis+credibility	0.7439	0.7598	0.2585	0.4288	0.3805	0.5423
DCLab2014_run3_VisTextClusterAvgRelevance	0.7927	0.7756	0.2578	0.4127	0.3838	0.5305
PeRCeiVe@UNICT2014_run2	0.7463	0.7553	0.2271	0.3902	0.3431	0.5063
BU-Retina_visdescSIFT_R5	0.7203	0.7228	0.2339	0.387	0.3492	0.4966
MIS2014_run3	0.6748	0.6732	0.2336	0.3985	0.3433	0.4949
Flickr initial results	0.8089	0.8065	0.2112	0.3427	0.3287	0.4699
DWS2014_run1_visualTestset(2)	0.7715	0.7524	0.2224	0.3405	0.3385	0.46
Miracl2014_run1_visualInformationOnly	0.8033	0.7772	0.2145	0.3265	0.3326	0.4501

Best improvements compared to Flickr (in percentage points): P@20 4.5, CR@20 13.

Results: Best Team Runs

*ranking based on official metrics (F1@20).

29

Results: Best Team Runs #2

*ranking based on official metrics (F1@20).

30
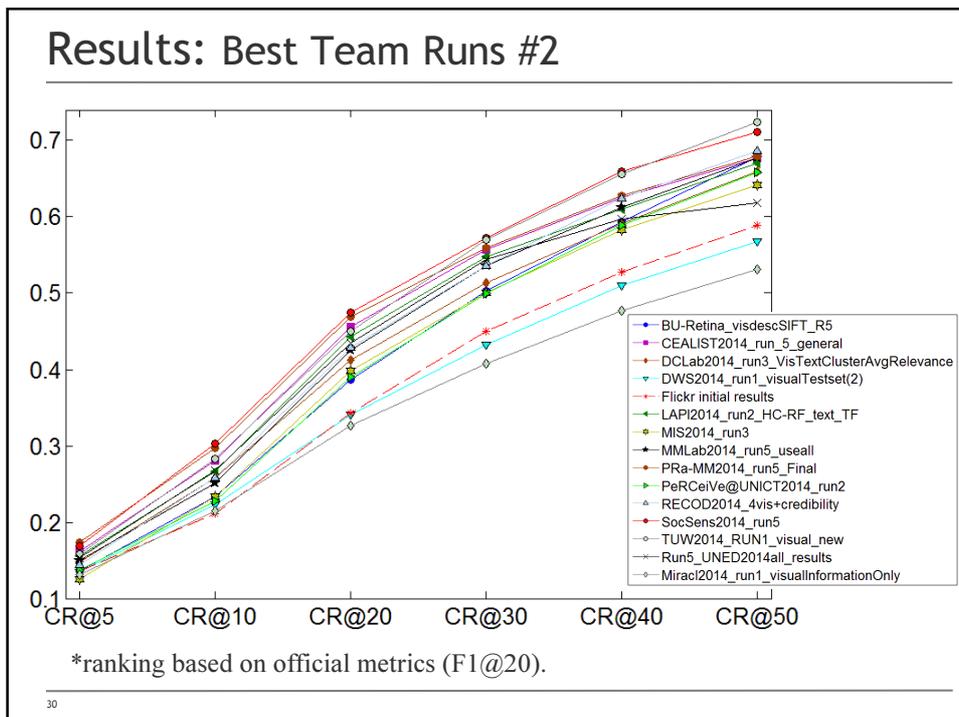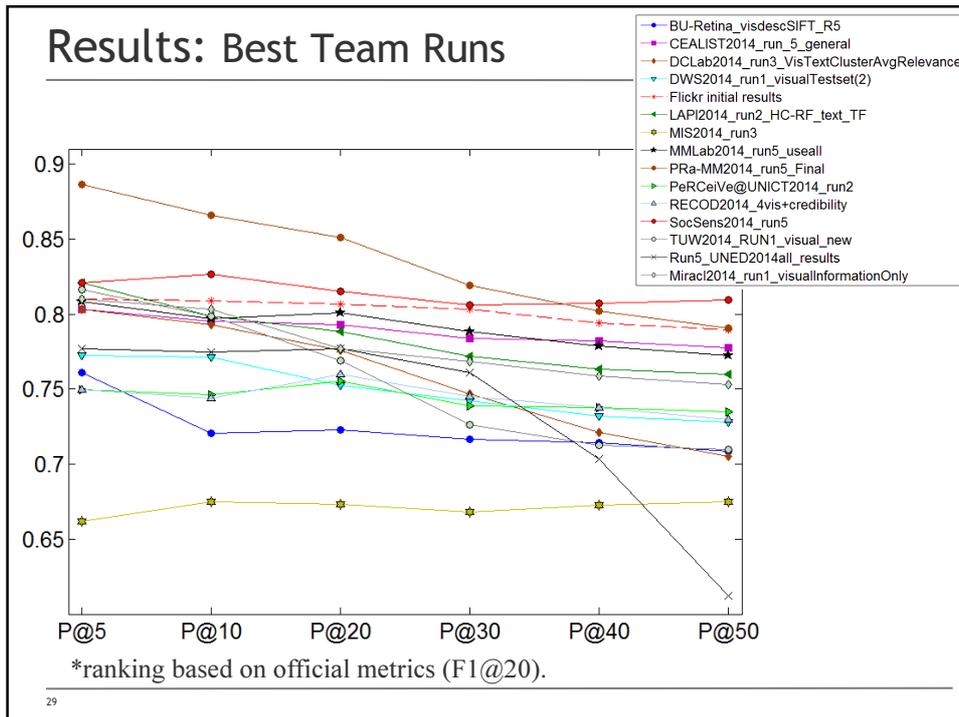
Results: Visual Results – Flickr Initial

Salisbury Cathedral
F1@20=0.3333,
P@20=1, CR@20=0.2,
25 clusters
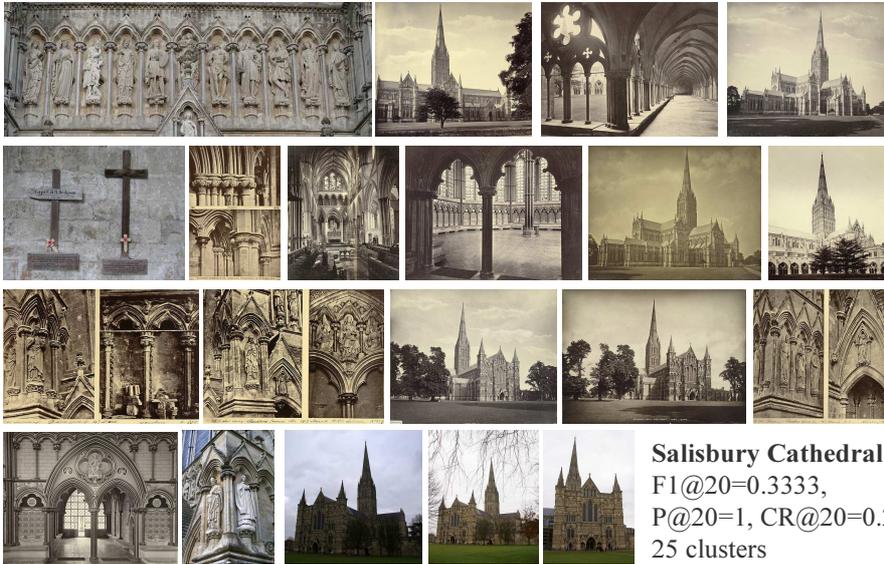
Results: Visual Results #2 –Best F1@20

Salisbury Cathedral
F1@20=0.6721,
P@20=0.95,
CR@20=0.52, 25 clust.
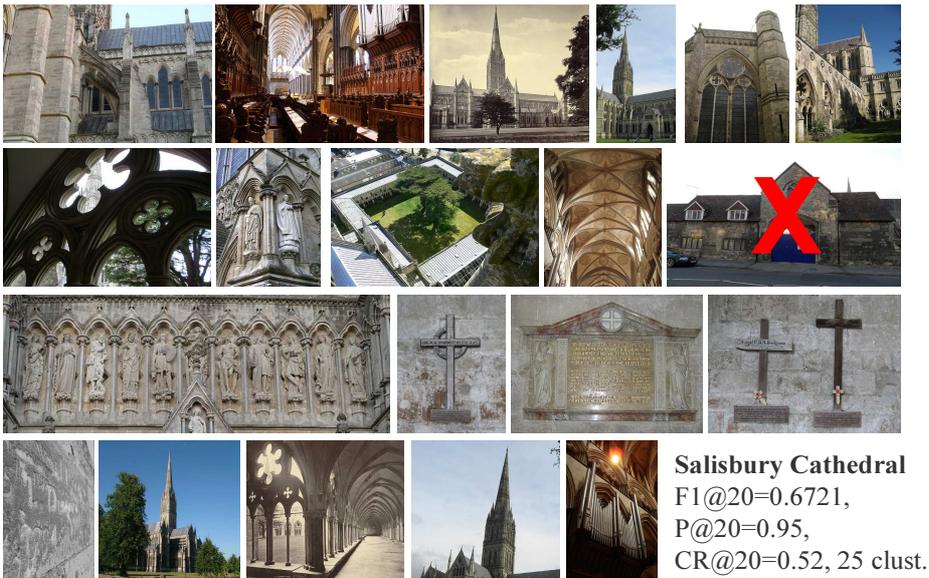
Results: Visual Results #3 – Lowest F1@20

Salisbury Cathedral
F1@20 = 0.3333,
P@20=1, CR@20=0.2,
25 clusters
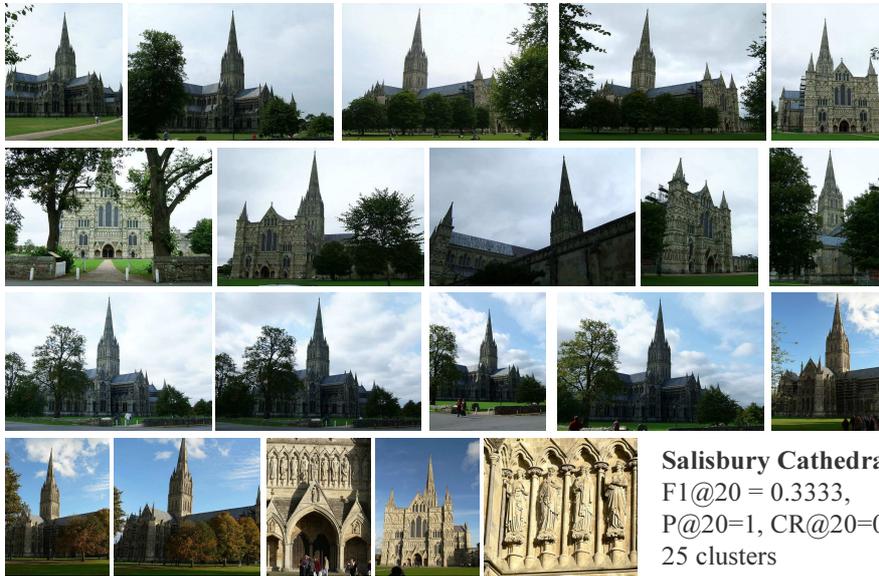
Brief Discussion

Methods:

▪ this year mainly clustering, re-ranking, optimization-based and relevance feedback (including machine-human);

▪ best run F1@20: pre-filtering + hierarchical clustering + tree refining + re-ranking using visual-text-cred. information (PRa-MM);

▪ user tagging credibility information proved its potential and should be further investigated in social retrieval scenarios.

Dataset:

▪ still low resources for location Creative Commons on Flickr;

▪ diversity annotation for 300 photos much difficult than for 100;

▪ descriptors were very well received (employed by most of the participants).

Present & Perspectives

For 2014:

- the task was a full task this year,

- the entire dataset is to be publicly released (soon).

For 2015:

- working on a new use case scenario.

35

Questions & Answers

Thank you!

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

See you at the poster session and for the technical retreat …

36