Overview of the Wikipedia Image Retrieval Task at ImageCLEF 2011

Theodora Tsikrika[1] and Adrian Popescu[2] and Jana Kludas[3]

[1] University of Applied Sciences Western Switzerland (HES-SO), Sierre, Switzerland
theodora.tsikrika@acm.org
[2] CEA, LIST, Vision & Content Engineering Laboratory, 92263 Fontenay aux Roses, France
adrian.popescu@cea.fr
[3] CUI, University of Geneva, Switzerland
jana.kludas@unige.ch

Abstract. ImageCLEF's Wikipedia Image Retrieval task provides a testbed for the system-oriented evaluation of multimedia and multilingual information retrieval from a collection of Wikipedia images. The aim is to investigate retrieval approaches in the context of a large and heterogeneous collection of images (similar to those encountered on the Web) that are searched for by users with diverse information needs. This paper presents an overview of the resources, topics, and assessments of the Wikipedia Image Retrieval task at ImageCLEF 2011, summarizes the retrieval approaches employed by the participating groups, and provides an analysis of the main evaluation results.

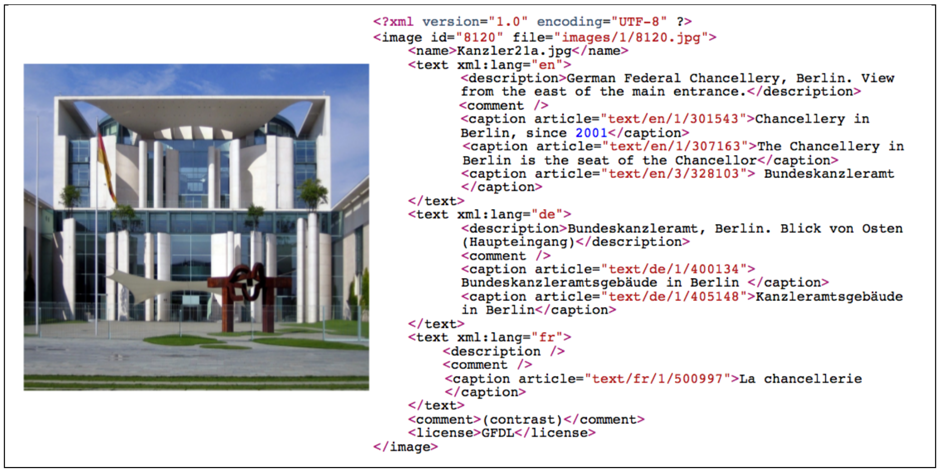
1 Introduction

The Wikipedia Image Retrieval task is an ad-hoc image retrieval task. The evaluation scenario is thereby similar to the classic TREC ad-hoc retrieval task: simulation of the situation in which a system knows the set of documents to be searched, but cannot anticipate the particular topic that will be investigated (i.e., topics are not known to the system in advance). Given a multimedia query that consists of a title in three different languages and a few example images describing a user's information need, the aim is to find as many relevant images as possible from a collection of Wikipedia images. Similarly to past years, participants are encouraged to develop approaches that combine the relevance of different media types and of multilingual textual resources into a single ranked list of results. A number of resources that support participants towards this research direction were provided this year.

The paper is organized as follows. First, we introduce the task's resources: the Wikipedia image collection and additional resources, the topics, and the assessments (Sections 2–4). Section 5 presents the approaches employed by the participating groups and Section 6 summarizes their main results. Section 7 concludes the paper.

2 Task resources

The ImageCLEF 2010 Wikipedia collection was used for the second time in 2011. It consists of 237,434 Wikipedia images, their user-provided annotations, the Wikipedia articles that contain these images, and low-level visual features of these images. The collection was built to cover similar topics in English, German, and French and it is based on the September 2009 Wikipedia dumps. Images are annotated in none, one or several languages and, wherever possible, the annotation language is given in the metadata file. The articles in which these images appear were extracted from the Wikipedia dumps and are provided as such. The collection is described in more detail in [11] and an example image with its associated metadata is given in Figure 1. A first set of image features were extracted using MM, CEA LIST's image indexing tool [7] and include both local (bags of visual words) and global features (texture, color and edges). An alternative set of global features, extracted with the MMRetrieval tool [13], was kindly provided by the Information Retrieval group at the Democritus University of Thrace, Greece (DUTH group).



```xml
<?xml version="1.0" encoding="UTF-8" ?>
<image id="8120" file="images/1/8120.jpg">
    <name>Kanzler21a.jpg</name>
    <text xml:lang="en">
          <description>German Federal Chancellery, Berlin. View
          from the east of the main entrance.</description>
          <comment />
          <caption article="text/en/1/301543">Chancellery in
          Berlin, since 2001</caption>
           <caption article="text/en/1/307163">The Chancellery in
           Berlin is the seat of the Chancellor</caption>
          <caption article="text/en/3/328103"> Bundeskanzleramt
          </caption>
    </text>
    <text xml:lang="de">
          <description>Bundeskanzleramt, Berlin. Blick von Osten
          (Haupteingang)</description>
          <comment />
          <caption article="text/de/1/400134">
          Bundeskanzleramtsgebäude in Berlin </caption>
          <caption article="text/de/1/405148">Kanzleramtsgebäude
          in Berlin</caption>
    </text>
    <text xml:lang="fr">
        <description />
        <comment />
        <caption article="text/fr/1/500997">La chancellerie
        </caption>
    </text>
    <comment>(contrast)</comment>
    <license>GFDL</license>
</image>
```

Fig. 1: Wikipedia image+metadata example from the ImageCLEF 2010 Wikipedia image collection.

3 Topics

The topics are descriptions of multimedia information needs that contain textual and visual hints.

3.1 Topic Format

These multimedia queries consist of a multilingual textual part, the query title, and a visual part made of several example images. The narrative of the query is only used during the assessment phase.

<**title**> query by keywords, one for each language: English, French, German
<**image**> query by image content (four or five)
<**narrative**> description of query in which an unambiguous definition of relevance and irrelevance is given

<title> The topic <*title xml:lang="en"*> has a language attribute that marks the English (en), French (fr) and German (de) topic title. It simulates a user who does not have (or want to use) example images or other visual constraints. The query expressed in the topic <title> is therefore a text-only query. This profile is likely to fit most users searching digital libraries or the Web.

Upon discovering that a text-only query does not produce many relevant hits, a user might decide to add visual hints and formulate a multimedia query.

<image> The visual hints are example images, which express the narrative of the topic.

<narrative> A clear and precise description of the information need is required in order to unambiguously determine whether or not a given document fulfils the given information need. In a test collection this description is known as the narrative. It is the only true and accurate interpretation of a user's needs. Precise recording of the narrative is important for scientific repeatability - there must exist, somewhere, a definitive description of what is and is not relevant to the user.

Textual terms and visual examples can be used in any combination in order to produce results. It is up to the systems how to use, combine or ignore this information; the relevance of a result does not directly depend on these constraints, but it is decided by manual assessments based on the <narrative>.

3.2 Topic Development

The 50 topics in the ImageCLEF 2011 Wikipedia Image Retrieval task (see Table 1), created by the organizers of the task, aim to cover diverse information needs and to have a variable degree of difficulty. They were chosen after a statistical analysis of a large scale image query log kindly provided by Exalead so as to cover a wide variety of topics commonly searched on the Web. Candidate topics were run through the Cross Modal Search Engine [4] (CMSE developed by

[4] `http://dolphin.unige.ch/cmse/`

the University of Geneva) in order to get an indication of the number of relevant images in top results for visual, textual and multimodal candidate queries.

The topics range from simple, and thus relatively easy (e.g., "brown bear"), to semantic, and hence highly difficult (e.g., "model train scenery"), with a balanced distribution of the two types of topics. One difference with 2010 [11] is the higher number of topics with named entities (and particularly known person names and products) proposed this year. This change is motivated by the results of the log analysis which confirmed that a lot of named entities are used in Web queries. Semantic topics typically have a complex set of constraints, need world knowledge, and/or contain ambiguous terms, so they are expected to be challenging for current state-of-the-art retrieval algorithms. We encouraged the participants to use multimodal and multilingual approaches since they are more appropriate for dealing with semantic information needs.

Image examples were selected from Flickr, after ensuring that they were uploaded under Creative Commons licenses. Each topic has four or five image examples, chosen so as to illustrate, to the extent possible, the visual diversity of the topic. Compared to 2010 [11], a larger number of images was provided per topic in order to have an improved visual characterization of the topics and thus to encourage multimodal approaches. Query image examples and their low-level features were also provided with the collection in order to ensure repeatability of the experiments. On average, the 50 topics contain 4.84 images and 3.1 words in their English formulation.

Table 1: Topics for the ImageCLEF 2011 Wikipedia Image Retrieval task: IDs, titles, the number of image examples providing additional visual information, and the number of relevant images in the collection.

ID	Topic title	# image examples	# relevant images
71	colored Volkswagen beetles	5	50
72	skeleton of dinosaur	5	116
73	graffiti street art on walls	5	95
74	white ballet dress	5	49
75	flock of sheep	5	34
76	playing cards	5	47
77	cola bottles or cans	5	24
78	kissing couple	5	33
79	heart shaped	5	34
80	wolf close up	4	25
81	golf player on green	5	22
82	model train scenery	5	40
83	red or black mini cooper	5	10
84	Sagrada Familia in Barcelona	5	7
85	Beijing bird nest	5	12
86	KISS live	5	11

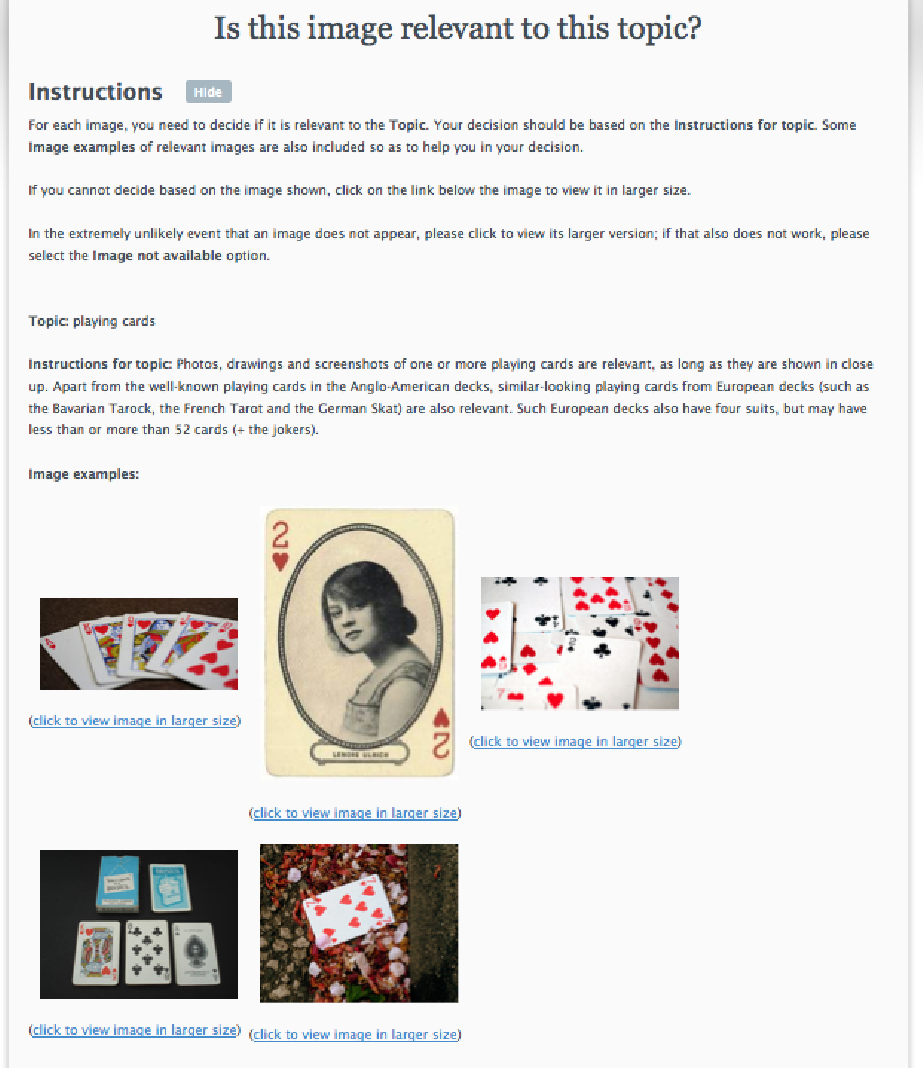
Continued on next page

Table 1 – continued from previous page

ID	Topic title	# image examples	# relevant images
87	boxing match	5	45
88	portrait of Segolene Royal	5	10
89	Elvis Presley	4	7
90	gondola in Venice	5	62
91	freestyle jumps with bmx or motor bike	5	18
92	air race	5	12
93	cable car	5	47
94	roller coaster wide shot	5	155
95	photo of real butterflies	5	112
96	shake hands	5	77
97	round cakes	5	43
98	illustrations of Alice's adventures in Wonderland	4	21
99	drawings of skeletons	5	95
100	brown bear	5	46
101	fountain with jet of water in daylight	5	141
102	black cat	5	20
103	dragon relief or sculpture	5	41
104	portrait of Che Guevara	4	13
105	chinese characters	5	316
106	family tree	5	76
107	sunflower close up	5	13
108	carnival in Rio	5	37
109	snowshoe hiking	4	12
110	male color portrait	5	596
111	two euro coins	5	58
112	yellow flames	5	92
113	map of Europe	5	267
114	diver underwater	5	33
115	flying bird	5	115
116	houses in mountains	5	105
117	red roses	4	27
118	flag of UK	4	12
119	satellite image of desert	4	93
120	bar codes	4	14

4 Assessments

The Wikipedia Image Retrieval task is an image retrieval task, where an image is either relevant or not (binary relevance). We adopted TREC-style pooling of



Is this image relevant to this topic?

Instructions

For each image, you need to decide if it is relevant to the **Topic**. Your decision should be based on the **Instructions for topic**. Some **Image examples** of relevant images are also included so as to help you in your decision.

If you cannot decide based on the image shown, click on the link below the image to view it in larger size.

In the extremely unlikely event that an image does not appear, please click to view its larger version; if that also does not work, please select the **Image not available** option.

Topic: playing cards

Instructions for topic: Photos, drawings and screenshots of one or more playing cards are relevant, as long as they are shown in close up. Apart from the well-known playing cards in the Anglo-American decks, similar-looking playing cards from European decks (such as the Bavarian Tarock, the French Tarot and the German Skat) are also relevant. Such European decks also have four suits, but may have less than or more than 52 cards (+ the jokers).

Image examples:

(click to view image in larger size)

(click to view image in larger size)

(click to view image in larger size)

(click to view image in larger size)

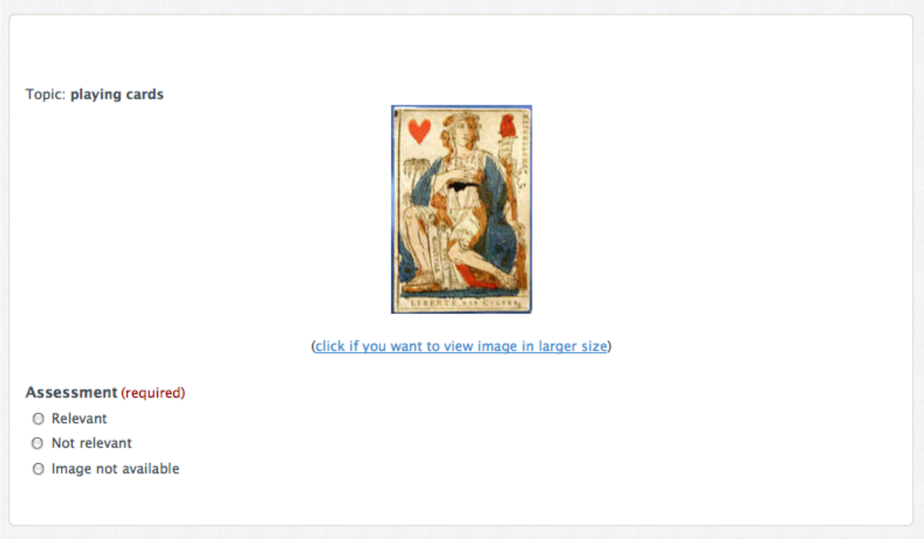
(click to view image in larger size)

Fig. 2: Instructions to workers for performing the relevance assessments.

the retrieved images with a pool depth of 100, resulting in pool sizes of between 764 and 2327 images with a mean of 1467 and median of 1440.

The relevance assessments were performed with a crowdsourcing approach using CrowdFlower (`http://crowdflower.com/`), a general-purpose platform for managing crowdsourcing tasks and ensuring high-quality responses. CrowdFlower enables the processing of large amounts of data in a short period of time by breaking a repetitive "job" into many "assignments", each consisting of small numbers of "units", and distributing them to many "workers" simultaneously. In our case, a *job* corresponded to performing the relevance assessments of the pooled images for a single topic, each *unit* was the image to be assessed, whereas each *assignment* consisted of assessing the relevance for a set of five units (i.e., images) for a single topic. The assessments were carried out by Amazon Mechanical Turk (`http://www.mturk.com`) *workers* based in the UK and the USA and each assignment was rewarded with $0.04$$.

For each assignment, each worker was provided with instructions in English for the English version of the topic, as shown in Figure 2 for topic 76, followed by five units to be assessed for that topic, each similar to the one shown in Figure 3. To prevent spammers and thus obtain accurate results, each assignment contained one "gold standard" image among the five images, i.e., an image already correctly labelled by the organizers. These gold standard data were used for estimating the workers' accuracy: if a worker's accuracy dropped below a threshold (70%), his/her assessments were excluded.



Topic: **playing cards**

(click if you want to view image in larger size)

Assessment (required)

○ Relevant

○ Not relevant

○ Image not available

Fig. 3: An image to be assessed.

For each topic, the gold standard data were created as follows. First, the images in the pool of depth 5 for that topic which could be unambiguously marked as relevant or non-relevant were assessed. This subset was selected so as to ensure that at least some relevant images were included in the gold standard set. If at the end of this round of assessment, the gold standard set contained less than 6% of the total images to be assessed for that topic, then further images from the original pool of depth 100 were randomly selected and assessed until the 6% limit was reached.

Each image was assessed by three workers with the final assessment obtained through a majority vote.

5 Participants

A total of 11 groups submitted 110 runs. The participation has slightly reduced compared to last year both in terms of number of participants (11 vs. 13) and of submitted runs (110 vs. 127). Nine participating groups out of 11 are located in Europe, one comes from Turkey and another one from Tunisia.

Table 2: Types of the 110 submitted runs.

Run type	**# runs**
Text (TXT)	51
Visual (IMG)	2
Text/Visual (TXTIMG)	57
Query Expansion (QE)	16
Relevance Feedback (RF)	15
QE & RF	12

Table 3: Annotation and topic language combinations in the textual and multimodal runs.

	Annotation language (AL)				
Topic Language (TL)	EN	DE	FR	EN+DE+FR	
EN	37	1	0	0	38
DE	0	2	0	0	2
FR	0	0	3	0	3
EN+DE+FR	0	0	0	65	65
	37	3	3	65	108

Table 2 gives an overview of the types of the submitted runs. Similarly to last year, more multimodal (text/visual) than text-only runs were submitted. Table 3 presents the combinations of annotation and topic languages used by participants in their textual and multimodal runs. The majority of submitted runs are multilingual in at least one of the two aspects. Most teams used both multilingual queries and multilingual annotations in order to maximize retrieval performance and the best results presented in the next section (see Tables 4 and 5) validate this approach. Although runs that implicate English only queries are by far more frequent than runs implicating German and French only, some participants also submitted the latter type of runs. A short description of the participants' approaches follows.

CEA LIST (9 runs - 5 single + 4 CEA-XRCE) [5] Their approach is mainly based on query expansion with Wikipedia. Given a topic, related concepts are retrieved from Wikipedia and used to expand the initial query. Then results are re-ranked using query models extracted from Flickr. They also used visual concepts (face/no face; indoor/outdoor) to characterize topics in terms of presence of these concepts in the image examples and to re-rank the results accordingly. Some of the runs submitted by CEA LIST (noted CEA-XRCE) were created using a late fusion of results with visual results produced by XRCE.

DBISForMaT (12 runs) [14] They introduced a retrieval model based on the polyrepresentation of documents which assumes that different modalities of a document can be combined in a structured manner to reflect a user's information need. Global image features were extracted using LIRE, a CBIR engine built on top of LUCENE. As it is underlined in [14], although promising, their results are hampered by the use of a naive textual representation of the documents.

DEMIR (6 runs) [3] They used the Terrier IR platform to test a large number of classical weighting schemes (BM25, TF-IDF, PL2 etc.) over a bag-of-words representation of the collection for text retrieval. They also performed a comparison of the visual descriptors provided by DUTH and report that the best purely visual results are obtained using the CEDD descriptor. Their multimodal runs are based on a late fusion approach and results show that merging modalities achieves small improvements compared to the textual results.

DUTH (19 runs) [1] The group has further developed its MMRetrieval engine that they introduced in 2010. It includes a flexible indexing of text and visual modalities as well as different fusion strategies (score combination and score normalization). This year, they introduced an estimation of query difficulty whose combination with score combination gave the best results. The group also kindly provided a set of low-level features which were used by a large number of participants.

ReDCAD (4 runs) [2] They focused on text retrieval and tested the use of the metadata related to the images as well as of the larger textual context of the images. LUCENE was used for both indexing and retrieving documents. Taking into account the textual context of the image is more effective than the use of the metadata only and a combination of the two provides a small additional improvement of results.

SINAI (6 runs) [8] The group submitted only textual runs and focused on an automatic translation of image descriptions from French and German to English. All their runs work with English queries only. Different linear combinations of image captions and descriptions were tested and they also combined results from Lemur and LUCENE retrieval engines. The combination of the two achieved the best results.

SZTAKI (10 runs) [6] The team used a retrieval system based on Okapi BM25 and also added synonyms from WordNet to expand the initial queries. Light Fisher vectors were used to represent low-level image features and then used to re-rank the top results obtained with purely textual retrieval. This late fusion procedure resulted in a slight degradation of performance compared to the textual run.

UAIC (6 runs) [4] For textual retrieval, they used the standard LUCENE search engine library and expanded some of the queries using WordNet synonyms. The visual search was performed using the Color and Edge Directionality Descriptor (CEDD) provided by the DUTH team. A linear combination of text and image results was performed which gave the best result.

UNED (20 runs) [9] They performed textual retrieval with a combination of IDRA, their in-house retrieval tool, and LUCENE and experimented with different settings (such as named entity recognition or use of Wikipedia articles). For multilingual textual runs, UNED tested early and late fusion strategies and the results show that the latter approach gives better results. Content based retrieval based on the CEDD features provided by DUTH was applied to the textual results. UNED tested both early and late fusion approached to obtain merged runs. Their fusion approaches were effective and the best results were obtained with a logistic regression feedback algorithm.

UNTESU (7 runs) [12] They applied Salient Semantic Analysis in order to expand queries with semantically similar terms from Wikipedia. A picturability measure was defined in order to boost the weight of terms which are associated to the initial topic in Flickr annotations. French and German annotations in the collection were translated to English and only English topics were used for the experiments. The best results were obtained with a combination of terms from the initial query and of expanded terms found using Lavrenko's relevance model.

XRCE (11 runs - 4 single + 7 XRCE-CEA) [5] For text retrieval, they implemented an information-based model and a lexical entailment IR model. Image content was described using spatial pyramids of Fisher Vectors and local RGB statistics. Late Semantic Combination (LSC) was exploited to combine results from text and image modalities. They showed that, although text retrieval largely outperforms pure visual retrieval, an appropriate combination of the two modalities results in a significant improvement over each modality considered independently. A part of the runs submitted by XRCE (noted XRCE-CEA) were created using a LSC approach which combined their text and visual runs as well as textual runs proposed by CEA LIST.

6 Results

Tables 4 and 5 present the evaluation results for the 15 best performing runs and the best performing run for each participant, respectively, ranked by Mean Average Precision (MAP). Similarly to 2010, the best runs were multimodal and multilingual. The best MAP performance (0.388) was reported for a run which combines XRCE and CEA LIST runs. The best textual run, ranked 11th, was also a combination of results from XRCE and CEA LIST and had a MAP of 0.3141. The results in Table 5 show that the best submitted runs were multimodal for eight out of nine participating groups that submitted such runs.

Table 4: Results for the top 15 runs.

Rank	**Participant**	**Run**	**Modality**	**FB/QE**	**AL**	**TL**	**MAP**	**P@10**	**P@20**	**R-prec.**
1	XRCE-CEA	SFLAXvis	Mix	FBQE	ENFRDE	ENFRDE	0.3880	0.6320	0.5100	0.4162
2	XRCE-CEA	AXmixFVSFL	Mix	FBQE	ENFRDE	ENFRDE	0.3869	0.6240	0.5030	0.4174
3	XRCE-CEA	SPLAXmixFVSFL	Mix	FBQE	ENFRDE	ENFRDE	0.3848	0.6200	0.4990	0.4174
4	XRCE-CEA	XTInn10AXmix	Mix	FBQE	ENFRDE	ENFRDE	0.3560	0.5340	0.4710	0.3835
5	XRCE	AXFVSFL	Mix	QE	ENFRDE	ENFRDE	0.3557	0.5940	0.4870	0.4051
6	XRCE	SPLAXFVSFL	Mix	FBQE	ENFRDE	ENFRDE	0.3556	0.5780	0.4840	0.4006
7	XRCE-CEA	SFLAXvis	Mix	FBQE	ENFRDE	ENFRDE	0.3471	0.5740	0.4450	0.3756
8	UNED	UNEDUV18	Mix	FB	ENFRDE	ENFRDE	0.3405	0.5420	0.4500	0.3752
9	UNED	UNEDUV20	Mix	FB	ENFRDE	ENFRDE	0.3367	0.5460	0.4410	0.3673
10	UNED	UNED-UV19	Mix	FB	ENFRDE	ENFRDE	0.3233	0.5400	0.4230	0.3586
11	XRCE-CEA	SPLAXmix	Txt	FBQE	ENFRDE	ENFRDE	0.3141	0.5160	0.4270	0.3504
12	XRCE-CEA	AXmix	Txt	QE	ENFRDE	ENFRDE	0.3130	0.5300	0.4250	0.3560
13	CEA LIST	mixFVSFL	Mix	QE	ENFRDE	ENFRDE	0.3075	0.5420	0.4210	0.3486
14	UNED	UNEDUV8	Txt	NOFB	ENFRDE	ENFRDE	0.3044	0.5060	0.4040	0.3435
15	UNED	UNEDUV14	Mix	FB	ENFRDE	ENFRDE	0.3006	0.5200	0.4030	0.3379

Table 5: Best performing run for each participant.

Rank	**Participant**	**Run**	**Modality**	**FB/QE**	**AL**	**TL**	**MAP**	**P@10**	**P@20**	**R-prec.**
1	XRCE-CEA	SFLAXvis	Mix	FBQE	ENFRDE	ENFRDE	0.3880	0.6320	0.5100	0.4162
8	UNED	UNEDUV18	Mix	FB	ENFRDE	ENFRDE	0.3405	0.5420	0.4500	0.3752
13	CEA-XRCE	mixFVSFL	Mix	QE	ENFRDE	ENFRDE	0.3075	0.5420	0.4210	0.3486
18	DUTH	QDSumw60	Mix	NOFB	ENFRDE	ENFRDE	0.2886	0.4860	0.3870	0.3401
23	UNTESU	BLRF	Txt	FB	EN	EN	0.2866	0.4220	0.3650	0.3276
53	DEMIR	Mix2	Mix	NOFB	ENFRDE	ENFRDE	0.2432	0.4520	0.3420	0.3001
61	ReDCAD	redcad02tx	Txt	NOFB	ENFRDE	ENFRDE	0.2306	0.3700	0.3060	0.2862
64	DBISForMaT	COMBINEDSW	Mix	NOFB	EN	EN	0.2195	0.4180	0.3630	0.2827
65	SZTAKI	txtjencolimg	Mix	FBQE	ENFRDE	ENFRDE	0.2167	0.4700	0.3690	0.2762
74	SINAI	lemurlucene	Txt	FB	EN	EN	0.2068	0.4020	0.3380	0.2587
94	UAIC2011	lucenecedd	Mix	NOFB	ENFRDE	ENFRDE	0.1665	0.4080	0.3090	0.2313

The complete list of results can be found at the ImageCLEF website [5].

[5] `http://www.imageclef.org/2011/wikimm-results`

6.1 Performance per modality for all topics

Here, we analyze the evaluation results using only the top 90% of the runs to exclude noisy and buggy results. Table 6 shows the average performance and standard deviation with respect to each modality. On average, the multimodal runs have better performance than textual ones with respect to all examined evaluation metrics (MAP, Precision at 20, and precision after R (= number of relevant) documents retrieved). This is in contrast with results reported in previous years when textual runs had better performances on average. This shift can be explained with changes in the resources as well as the approaches this year, i.e., increased number of visual examples in the queries, improved visual features and more appropriate fusion techniques used by the participants.

Table 6: Results per modality over all topics.

Modality	**MAP**		**P@20**		**R-prec.**	
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**
All top 90% runs (100 runs)	0.2492	0.056	0.3597	0.0577	0.2989	0.0498
Mix in top 90% runs (50 runs)	0.2795	0.0498	0.3940	0.0494	0.3289	0.0414
Txt in top 90% runs (50 runs)	0.2189	0.0445	0.3254	0.0432	0.2689	0.0381

6.2 Performance per topic and per modality

To analyze the average difficulty of the topics, we classify the topics based on the AP values per topic averaged over all runs as follows:

easy: $MAP > 0.3$
medium: $0.2 < MAP <= 0.3$
hard: $0.1 < MAP <= 0.2$
very hard: $MAP < 0.1$.

Table 7 presents the top 10 topics per class (i.e., easy, medium, hard, and very hard), together with the total number of topics per class. Out of 50 topics, 23 fall in the hard or very hard classes. This was actually intended during the topic development process, because we opted for highly semantic topics that are challenging for current retrieval approaches. 7 topics were very hard to solve($MAP < 0.10$). The topic set includes only 17 easy topics (such as "illustrations of Alice's adventures in Wonderland", "Sagrada Familia in Barcelona", "colored Volkswagen beetles", "KISS live"). Similarly to last year, a large number of the topics in the easy and medium classes include a reference to a named entity and, consequently, are easily retrieved with simple textual approaches. As for very hard topics, they often contain general terms ("cat", "house", "train" or "bird"), which have a difficult semantic interpretation or high concept variation and are, hence, very hard to solve.

Table 7: Topics classified based on their difficulty (up to 10 topics per class) - the total number of topics per class is given in the table header.

easy (17 topics)	**medium (12 topics)**	**hard (14 topics)**	**very hard (7 topics)**
98 illustrations of Alice's adventures in Wonderland	**108** carnival in Rio	**99** drawings of skeletons	**82** model train scenery
88 portrait of S. Royal	**92** air race	**117** red roses	**87** boxing match
84 Sagrada Familia in Barcelona	**120** bar codes	**101** fountain with jet of water in daylight	**115** flying bird
89 Elvis Presley	**91** freestyle jumps with bmx or motor bike	**93** cable car	**118** flag of UK
94 roller coaster wide shot	**104** portrait of Che Guevara	**80** wolf close up	**102** black cat
96 shake hands	**114** diver underwater	**112** yellow flames	**110** male color portrait
71 colored VW Beetles	**83** red or black mini cooper	**90** gondola in Venice	**116** houses in mountains
86 KISS live	**79** heart shaped	**78** kissing couple	
107 sunflower close up	**73** graffiti street art on walls	**95** photos of real butterflies	
77 cola bottles or cans	**106** family tree	**119** satellite image of desert	

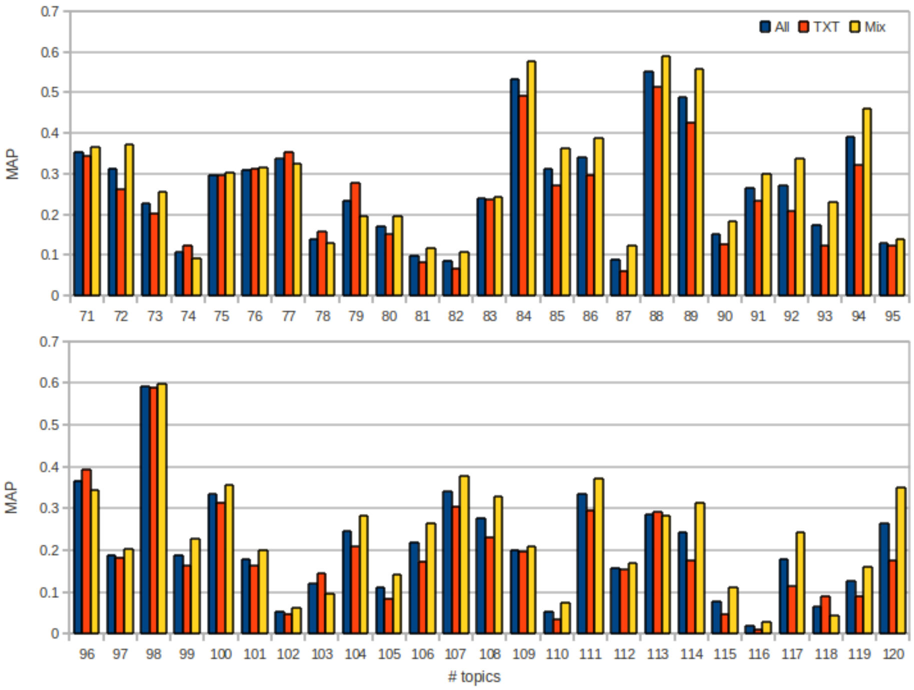


Fig. 4: Average topic performance over all, text-only, and mixed runs.

6.3 Visuality of topics

We also analyzed the performance of runs that use only text (TXT) versus runs that use both text and visual resources (MIX). Figure 4 shows the average performance on each topic for all, text-only and text-visual runs. The multimodal runs outperform the textual ones in 42 out of the 50 topics and the textual runs outperform mixed runs in 8 cases. This indicates that most of the topics benefit from a multimodal approach.

The "visuality" of topics can be deduced from the performance of text-only and text-visual approaches that were presented in the last section. We consider that, if for a topic the text-visual approaches improve significantly the MAP over all runs (i.e., by *diff*$(MAP) >= 0.01$), then we could consider that to be a visual topic. In the same way, we can define topics as textual, if the text-only approaches improve significantly the MAP over all runs of a topic. Based on this analysis, 38 of the topics can be characterized as visual and 7 as textual. The remaining 5 topics, where no clear improvements are observed, are considered to be neutral. Compared to 2010, when there were more textual than visual topics, the distribution of topics in visual vs. textual changed significantly. As with the aggregate run performances, this change is most probably a result of the increased number of query images, the improved low-level image indexing as well as the better fusion techniques proposed this year.

Table 8 presents the topics in each group, as well as some statistics on the topic, their relevant documents, and their distribution over the classes that indicate their difficulty. Given that there are only few textual and neutral topics, it is difficult to provide a robust analysis of the characteristics of the topics of each type.

Table 8: Best performing topics for textual and text-visual runs relative to the average over all runs (up to 10 topics per type).

	textual (7 topics)	**visual (38 topics)**	**neutral (5 topics)**
Topics	**79** heart shaped	**120** bar codes	**98** illustrations of Alice's adventures in Wonderland
	96 shake hands	**94** roller coaster wide shot	**75** flock of sheep
	118 flag of UK	**114** diver underwater	**83** red or black mini cooper
	103 dragon relief of sculpture	**89** Elvis Presley	**76** playing cards
	74 white ballet dress	**92** air race	**113** map of Europe
	77 cola bottles or cans	**72** skeleton of dinosaur	
	78 kissing couple	**93** cable car	
		108 carnival in Rio	
	106 family tree		
		85 Beijing bird nest	
#words/topic	2.857	3.026	3.8
#reldocs	38.57	73.44	75.8
MAP	0.238	0.244	0.369
easy	2	11	4
medium	1	10	1
hard	3	11	0
very hard	1	6	0

The number of words per topic is larger for neutral queries than for textual and visual ones. The average number of relevant documents is significantly smaller for textual topics compared to the other two classes whereas the average MAP is bigger for neutral topics.

The distribution of the textual, visual and neutral topics over the classes expressing their difficulty shows that the visual and textual topics are more likely to fall into the hard/very hard class than the neutral ones.

6.4 Effect of Query Expansion and Relevance Feedback

Finally, we analyze the effect of the application of query expansion (QE), relevance feedback (FB) techniques as well as of their combination (FBQE). Similarly to the analysis in the previous section, we consider the techniques to be useful for a topic, if they improved significantly the MAP over all runs. Table 9 presents the best performing topics for these techniques and some statistics. Query expansion is useful only for 3 topics and relevance feedback for 10. Interestingly, a combination of query expansion and of relevance feedback is effective for a much larger number of topics (33 out of 50). Expansion and feedback tend to be more useful for topics that are either hard or very hard compared to easy or medium topics.

Table 9: Best performing topics for query expansion (QE) and feedback (FB) runs relative to the average over all runs. Only the top 10 topics that benefit from query expansion are presented here.

	QE (3 topics)	**FB (10 topics)**	**FBQE (33 topics)**
Topics	**118** flag of UK	**99** drawings of skeletons	**80** wolf close up
	80 wolf close up	**79** heart shaped	**76** playing cards
	90 gondola in Venice	**95** photo of real butterflies	**74** white ballet dress
		93 cable car	**97** round cakes
		115 flying bird	**91** freestyle jumps with bmx or motor bike
		78 kissing couple	**120** bar codes
		74 white ballet dress	**114** diver underwater
		92 air race	**96** shake hands
		73 graffiti street art on walls	**101** fountain with jet of water in daylight
		82 model train scenery	**94** roller coaster wide shot
#words/topic	3	2.8	3.273
#reldocs	33	63.2	79.848
avg. MAP	0.156	0.202	0.2153
easy	0	0	8
medium	0	3	8
hard	2	5	12
very hard	1	2	5

7 Conclusions

For the second time this year, a multimodal and multilingual approach performed best in the Wikipedia Image Retrieval task. The majority of runs focused either on a combination of topic languages or on English queries only,

only a few runs were submitted for German and French queries only. Multilingual runs perform clearly better than monolingual ones due to the distribution of the information over the different languages.

It is encouraging to see that more than half of the submitted runs were multimodal and that the best submitted runs were multimodal for eight out of nine participating groups that submitted such runs. Many of the participants in the Wikipedia Image Retrieval Task have participated in the past and thus have been able to improve their multimodal retrieval approaches continuously. For the first time this year, there was a cooperation of two of the participating groups for testing late fusion of their results which is an interesting development.

A further analysis of the results showed that most topics (42 out of 50) were significantly better solved with multimodal approaches. This is not only due to the improvement of the fusion approaches mentioned above, but also due to an increased number of query images compared to the last years and improved visual features. Finally, we found that expansion and feedback techniques tend to be more useful for topics that are either hard or very hard compared to easy or medium topics.

8 Acknowledgements

Theodora Tsikrika was supported by the EU in the context of Promise (contract no. 258191) and Chorus+ (contract no. 249008) FP7 projects. Adrian Popescu was supported by the French ANR (Agence Nationale de la Recherche) via the Georama project (ANR-08-CORD-009). Jana Kludas was funded by the Swiss National Fund (SNF).

The authors would also like to thank the Information Retrieval group at the Democritus University of Thrace, Greece (DUTH group) for sharing visual features with all other task participants.

References

1. Avi Arampatzis, Konstantinos Zagoris, and Savvas A. Chatzichristofis. DUTH at ImageCLEF 2011 Wikipedia Retrieval. In Petras et al. [10].
2. Hatem Awadi, Mouna Torjmen Khemakhem, and Maher Ben Jemaa. Evaluating some contextual factors for image retrieval: ReDCAD participation at ImageCLEFWikipedia 2011. In Petras et al. [10].
3. Tolga Berber, Ali Hosseinzadeh Vahid, Okan Ozturkmenoglu, Roghaiyeh Gachpaz Hamed, and Adil Alpkocak. DEMIR at ImageCLEFwiki 2011: Evaluating Different Weighting Schemes in Information Retrieval. In Petras et al. [10].
4. Emanuela Boros, Alexandru-Lucian Ginsca, and Adrian Iftene. UAIC's participation at Wikipedia Retrieval @ ImageCLEF 2011. In Petras et al. [10].
5. Gabriela Csurka, Stéphane Clinchant, and Adrian Popescu. XRCE and CEA LIST's Participation at Wikipedia Retrieval of ImageCLEF 2011. In Petras et al. [10].
6. Bálint Daróczy, Róbert Pethes, and András A. Benczúr. SZTAKI @ ImageCLEF 2011. In Petras et al. [10].

7. Bertrand Delezoide, Hervé Le Borgne, Romaric Besançon, Gaël De Chalendar, Olivier Ferret, Faïza Gara, Patrick Hède, Meriama Laib, Olivier Mesnard, Pierre-Alain Moellic, and Nasredine Semmar. MM: modular architecture for multimedia information retrieval. In *Proceedings of the 8th International Workshop on Content-Based Multimedia Indexing (CBMI 2010)*, pages 136–141, 2010.
8. Miguel Ángel García-Cumbreras, Manuel Carlos Díaz-Galiano, L. Alfonso Ureña-López, and Javier Arias-Buendía. SINAI at ImageCLEF Wikipedia Retrieval task 2011: testing combined systems. In Petras et al. [10].
9. Ruben Granados, Joan Benavent, Xaro Benavent, Esther de Ves, and Ana García-Serrano. Multimodal information approaches for the Wikipedia collection at ImageCLEF 2011. In Petras et al. [10].
10. Vivien Petras, Pamela Forner, and Paul Clough, editors. *CLEF 2011 working notes*, 2011.
11. Adrian Popescu, Theodora Tsikrika, and Jana Kludas. Overview of the wikipedia retrieval task at imageclef 2010. In *Working notes of ImageCLEF 2010*, 2010.
12. Miguel E. Ruiz, Chee Wee Leong, and Samer Hassan. UNT at ImageCLEF 2011: Relevance Models and Salient Semantic Analysis for Image Retrieval. In Petras et al. [10].
13. Konstantinos Zagoris, Avi Arampatzis, and Savvas A. Chatzichristofis. www.mmretrieval.net: a multimodal search engine. In *Proceedings of the Third International Conference on SImilarity Search and APplications*, SISAP '10, pages 117–118, New York, NY, USA, 2010. ACM.
14. David Zellhöfer and Thomas Böttcher. BTU DBIS' Multimodal Wikipedia Retrieval Runs at ImageCLEF 2011. In Petras et al. [10].