Overview of the Wikipedia Retrieval Task at ImageCLEF 2010

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

[1] Institut Télécom/Télécom Bretagne, Brest, France
adrian.popescu@telecom-bretagne.eu
[2] CWI, Amsterdam, The Netherlands
theodora.tsikrika@acm.org
[3] CUI, University of Geneva, Switzerland
jana.kludas@unige.ch

Abstract. ImageCLEF's Wikipedia Retrieval task provides a testbed for the system-oriented evaluation of multimedia 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 Retrieval task at ImageCLEF 2010, summarizes the retrieval approaches employed by the participating groups, and provides an analysis of the main evaluation results.

1 Introduction

The Wikipedia 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 and one or more example images describing a user's multimedia information need, the aim is to find as many relevant images as possible from a Wikipedia image collection.

The Wikipedia Retrieval task differs from other benchmarks in multimedia information retrieval, like TRECVID, in the sense that the textual modality in the Wikipedia image collection contains less noise than the speech transcripts in TRECVID. Similarly to past years, participants are encouraged to develop approaches that combine the relevance of different media types 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 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. Image features were extracted using MM, CEA LISTs image indexing tool [6] and include both local (bags of visual words) and global features (texture, color and edges).

The main difference between the ImageCLEF 2010 Wikipedia collection and the INEX MM Wikipedia collection [19] used in the ImageCLEF WikipediaMM 2008-2009 tasks is that multilinguality has been added and both mono- and cross-lingual evaluations can be carried out. Another difference is that participants received for each image both its user-provided annotations, similarly to before, but also links to the article(s) which contain the image.

The collection contains 237,434 images and the associated annotations are distributed as follows:

- English only: 70,127
- German only: 50,291
- French only: 28,461
- English and German: 26,880
- English and French: 20,747
- German and French: 9,646
- English, German and French: 22,899
- Language undetermined: 8,144
- No textual annotation: 239

This distribution shows that the annotations in the ImageCLEF 2010 Wikipedia collection are heterogeneous, with nearly 10% of the collection with annotations in all three languages, 24% of the images with annotations in two languages out of three, 62% of the images with annotations in only one language, and the rest of the images with annotations for which language was not identified or no annotation exists. This distribution of annotations aims to encourage the investigation of multilingual approaches since they are likely to work better than monolingual approaches.

2.1 Metadata

Metadata are provided as a single metadata.zip archive which is split into 26 directories (from 1 to 26): "metadata/1" contains XML files from 0.xml to 9999.xml, "metadata/2" contains images from 10000.xml to 19999.xml, etc. Note that each directory may contain the metadata for less than 10,000 images, since some

of the initial images were removed during the collection construction so as to eliminate duplicates and to ensure, to the extent possible, that the provided images are copyright free and valid. Textual annotations were extracted from the Wikimedia Commons files that describe the images and from the article(s) that contain the images. Annotations are grouped by language when it was possible to identify their language.

The main components of the .xml files (for an example see Figure 1) are:

- <image> - contains the unique ID of the image and a link to the image file.
- <name> - the name of the image as found in the Wikimedia Commons repository. No processing has been applied to this text.
- <text xml:lang="LANGUAGE"> - where LANGUAGE if one of {en, de, fr}. These are textual annotations for which the language was identified.
  - <description> was extracted from the Wikimedia Commons page of the image whenever the language of this text was explicitly marked in a normalized manner.
  - <comment> was extracted from the Wikimedia Commons page of the image whenever the language of this text was marked in a normalized manner. <comment> is a substring of the raw comment described below.
  - <caption> is the text that accompanies the image in Wikipedia articles which are provided in the TEXT part of the collection (see Section 2.3) and are linked in the <caption> element. Sometimes Wikipedia articles contain images without captions. In such cases the <caption> element links to the article but is empty. Note that one image can appear in more than one article; in that case, all captions or links to the articles are provided.
- <comment> - raw annotation as found on the Wikimedia Commons page of the image. No processing was applied to this text and the comment can be in one or more languages (not necessarily one of {en, de, fr}).
- <license> - licensing information extracted from the Wikimedia Commons page of the image.

Any or all of the textual annotations of the images can be missing and some of them can be redundant.

2.2 Images

The images are provided in the "images" directory and, due to their total size, are provided in 26 separate archives. They are organized in the same way as the associated metadata, namely in 26 directories and have the same IDs as the XML files (images/1/7924.png corresponds to metadata/1/7924.xml). Images are linked in the <image> element of the XML files using "file".

```
<?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.

2.3 Text

Wikipedia article texts are provided in the "text" directory and are separated in language subdirectories ("text/en", "text/de", "text/fr"). Articles were renamed in order to have unique IDs in the collection and their IDs start at 300000 for English, 400000 for German and 500000 for French. They are grouped in 5 subdirectories for each language, with each such subdirectory containing 10,000 elements or less (for instance, "text/en/1/" includes articles from 300000 to 309999). Articles in which images appear are referenced in the <caption> element of the XML files contained in the "metadata" directory.

2.4 Additional Resources

Image features were also provided to support the participants in their investigation of multimodal approaches.

Image features are provided in the "features" directory in binary files (one file per feature). Features were computed using the MM, CEA LIST's image indexing tool [6] and they include:

- cime (features/cime.txt) a border/interior classification algorithm proposed by [16] which classifies pixels into interior or border and then builds a 64 bins histogram for each pixel type. The feature space is composed of 64 dimensions.
- tlep (features/tlep.txt) a descriptor which combines image texture and color [3]. Two texture histograms are built for edge and non-edge pixels and a 64 bins color histogram compose the image description. The feature space is composed of 576 dimensions.

– bag (features/bag.txt) a descriptor based on bags of visual words [14]. A vocabulary containing 5000 visual words was built from a random sample of the collection and all the images were then indexed with elements of this vocabulary. The feature space is composed of 5000 dimensions.

Each line in the feature files is composed of the image id (first column of each line) and the representation of the image in the feature space (from second to the last column of each line). Columns are separated by white spaces.

Along with the feature files a Perl script was provided to show how to exploit features in order to compute image similarities.

The additional resources are beneficial to researchers who wish to exploit visual evidence without performing image analysis. Of course, participants could also extract their own image features.

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 textual part, the query title, and a visual part, one or several example images.

<**title**> query by keywords
<**image**> query by image content (one or several)
<**narrative**> description of query in which the definitive definition of relevance and irrelevance are given

<title> The topic <title> 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 Internet.

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 topics in the ImageCLEF 2010 Wikipedia 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 from an initial pool of 137 candidate topics that were derived from a search log file and from the topics of the 2008 and 2009 WikipediaMM tasks. Candidate topics were run through the Cross Modal Search Engine [4] (CMSE - developed by the University of Geneva) in order to get an indication of the number of relevant images in top results for baseline image only, text only and multimodal approaches. The final pool contains around 1/3 of topics that return good results for each type of retrieval.

The topics range from simple, and thus relatively easy (e.g., "postage stamp"), to semantic, and hence highly difficult (e.g., "white house with garden"), with the latter forming the bulk of the topics. 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 approaches since they are more appropriate for dealing with semantic information needs.

Image examples were selected from Flickr, after checking that they are uploaded under the Creative Commons license. Each topic has one or several image examples, chosen so as to illustrate the visual diversity of the topic. Query image examples and their low-level features are also associated to the collection in order to ensure repeatability of the experiments. On average, the 70 topics contain 1.68 images and 2.7 words.

Table 1: Topics for the ImageCLEF 2010 Wikipedia 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
1	fractals	2	317
2	cockpit of an airplane	1	87
3	basketball game close up	2	116
4	Christmas tree	2	22
5	Oktoberfest beer tent	2	9
6	solar panels	2	101
7	lightning in the sky	1	43

Continued on next page

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

Table 1 – continued from previous page

ID	Topic title	# image examples	# relevant images
8	tennis player on court	2	393
9	flying hot air balloon	2	30
10	horseman	2	96
11	landline telephone	1	27
12	DNA helix	1	39
13	trains and locomotives	2	687
14	videogames screenshot	2	114
15	cyclist	2	176
16	spider with cobweb	2	27
17	beach volleyball	2	7
18	stars and galaxies	2	384
19	lochs in Scotland	1	53
20	mountains with sky	2	969
21	Chernobyl disaster ruins	2	17
22	sharks underwater	2	27
23	emoticon smiley	2	8
24	Rorschach black and white	1	6
25	Shiva painting or sculpture	2	29
26	brain scan	2	24
27	active volcano with ash cloud	1	75
28	palm trees	2	71
29	desert scenery	2	247
30	harbour	2	454
31	yellow buses	1	50
32	people laughing	2	51
33	close up of antenna	2	90
34	people playing guitar	2	348
35	race car	2	852
36	portrait of Jintao Hu	1	5
37	close up of bottles	1	237
38	baseball game	1	140
39	cactus in desert	1	13
40	ferrari red	1	185
41	polar bear	2	46
42	Paintings related to cubism	2	23
43	skyscraper in daylight	2	362
44	saturn	2	81
45	snowy winter landscape	2	376
46	sailboat	1	181
47	soccer stadium	1	366
48	civil airplane	1	633

Continued on next page

Table 1 – continued from previous page

ID	Topic title	# image examples	# relevant images
49	surfing on waves	1	38
50	portraits of people	2	1727
51	aerial pictures of landscapes	3	678
52	satellite image	2	875
53	ISS international space station	1	178
54	launching space shuttle	1	102
55	building site	1	125
56	musician on stage	1	568
57	road street signs	2	305
58	red fruits	2	146
59	cities at night	3	528
60	notes on music sheet	1	233
61	earth from space	2	89
62	Shopping in a market	2	224
63	postage stamp	3	866
64	woman in red dress	2	57
65	sea sunset or sunrise	1	116
66	bridges in daylight	2	793
67	white house with garden	1	77
68	historic castle	2	605
69	red tomato	1	33
70	close up of trees	2	603

4 Assessments

The Wikipedia Retrieval task is an image retrieval task, where an image with its metadata is either relevant or not (binary relevance). We adopted TREC-style pooling of the retrieved images with a pool depth of 100, resulting in pool sizes of between 1421 and 3850 images with a mean of 2659 and median of 2531. The evaluation was performed by three participant groups and by the organizers within a period of 4 weeks after the submission of runs. The assessors used a modified version of the web-based interface that was used last year and which has also been previously employed in the INEX Multimedia and TREC Enterprise tracks.

5 Participants

A total of 13 groups submitted 127 runs. The participation was significantly higher than last year both in terms of number of participants (13 vs. 8) and of submitted runs (127 vs. 57). Although the highest number of groups are located in European countries, the geographic spread of participants has increased this year, with North-American and Asian groups being better represented.

Table 2: Types of the 127 submitted runs.

Run type	**# runs**
Text (TXT)	48
Visual (IMG)	7
Text/Visual (TXTIMG)	72
Query Expansion (QE)	18
Relevance Feedback (RF)	14
Pseudo RF	9
QE & RF	1

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

	Annotation language				
Query Language	EN	DE	FR	EN+DE+FR	
EN	32	0	0	13	45
DE	0	4	0	0	4
FR	0	0	5	1	6
EN+DE+FR	1	0	0	64	65
	33	4	5	78	120

Table 2 gives an overview of the types of the submitted runs. This year more multimodal (text/visual) than text-only runs were submitted. Table 3 presents the combinations of annotation and query languages used by participants in their textual and multimodal runs. A majority of submitted runs are multilingual in at least one of the two aspects. Many 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.

CHESHIRE (8 runs) [8] Their focus was on textual retrieval and they proposed runs for English, French and German queries. The retrieval model used is logistic regression, complemented with blind relevance feedback.

DAEDALUS (6 runs) [7] They proposed only textual runs and experimented with corpus and topic expansion using several collection metadata, but also information about named entities and concepts included in DBPedia.

DCU (3 runs) [9] Their approach was based on document expansion with Wikipedia content and using the Okapi feedback algorithm. In addition, document reduction was exploited to weight the terms in the query. Okapi BM25 was used for the retrieval phase and only English queries were examined.

DUTH (20 runs) [1] They experimented with the usage of all different modalities (textual descriptions in each language, image features) and discussed two types of fusion methods: score normalization and score combination.

They concluded that the text modality is by far more important than the image modality since the latter results only in little improvement when introduced into the retrieval framework.

I2RCVIU (6 runs) [18] They presented results for mono- and multilingual textual runs as well as for multimodal runs. One interesting result they report is the use of visual near duplicates in order to boost images that are very similar to top results from textual runs.

NUS (14 runs) They submitted both text and multimodal runs. For textual runs, they mapped image metadata to Wikipedia concepts, which were subsequently used during retrieval. For mixed runs, they also identify concepts from images.

RGU (8 runs) [11] They extended their quantum theory approach first presented at ImageCLEF 2007. A tensor product model is developed to represent textual and visual features in a non-separable composite system. They also introduced a new "bag of visual words" inspired image features.

SZTAKI (5 runs) [5] Their approach used Okapi BM25 text retrieval and Histogram of Oriented Gradients features clustered with Gaussian Mixture Models for image description. Query expansion with visual information was performed over textual results and this resulted in a slight improvement of the final results.

TELECOM (16 runs) [12] 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.

UAIC (2 runs) From the image's textual metadata, they generated image keywords which were filtered using a comparison to visually similar images. During retrieval, the topics were matched against the previously generated keywords. Visual similarity was equally used in order to rank retrieved images.

UNED (20 runs) [2] They implemented a variant of VSM approach with TF-IDF weights and used it for their best textual run. For the multimodal runs, they used late fusion with three different algorithms: automatic, query expansion and relevance feedback based on logistic regression.

UNT (3 runs) [13] Their main goal was to explore the use of cross-lingual information retrieval. Instead of using a standard automatic query technique, they translated textual metadata to the language of the query and then performed mono-lingual retrieval. They also used manual query expansion in order to add possibly useful words to the query. This interactive retrieval technique improves the precision of top results, but does not improve the overall performance of the system.

XRCE (16 runs) [4] They represented textual metadata using standard language models or a power law. Image content was described using Fisher Vectors improved with power and L2 normalization and spatial pyramid representations. 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.

6 Results

Tables 4 and 5 present the evaluation results for the 15 best performing runs and the best performing run for each team, respectively, ranked by Mean Average Precision (MAP). Compared to 2009, when the best submitted runs were textual, this year multimodal runs submitted by XRCE were ranked best with a MAP of 0.2765. The best textual run, also submitted by XRCE was ranked 12th and had a MAP of 0.2361. The results in Table 5 show that results for individual teams are more nuanced, with five teams having multimodal runs as their best submission.

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	XAFSQTAMP	Mixed	FB	EN+FR+DE	EN+FR+DE	0.2765	0.5814	0.5193	0.3465
2	xrce	XACFSRTAMP	Mixed	FB	EN+FR+DE	EN+FR+DE	0.2681	0.5686	0.5257	0.3413
3	xrce	XAFSRTAMPRT	Mixed	FB	EN+FR+DE	EN+FR+DE	0.2627	0.6114	0.5407	0.3289
4	xrce	XAFSRPTPAMP	Mixed	FB	EN+FR+DE	EN+FR+DE	0.2591	0.5829	0.5143	0.3316
5	xrce	XACFSRTAMP2	Mixed	NOFB	EN+FR+DE	EN+FR+DE	0.2575	0.5957	0.5164	0.3257
6	xrce	XACFSRTAMP3	Mixed	FB	EN+FR+DE	EN+FR+DE	0.2532	0.5429	0.4986	0.3300
7	xrce	XAFSRTAMPRT2	Mixed	FB	EN+FR+DE	EN+FR+DE	0.2527	0.5971	0.5336	0.3200
8	xrce	XAFSRPTPAMP2	Mixed	NOFB	EN+FR+DE	EN+FR+DE	0.2495	0.5729	0.5157	0.3189
9	xrce	XAFSRTAMPRT3	Mixed	FB	EN+FR+DE	EN+FR+DE	0.2493	0.5171	0.4743	0.3233
10	xrce	XACFSRTAMP2	Mixed	NOFB	EN+FR+DE	EN+FR+DE	0.2424	0.5543	0.4907	0.3183
11	xrce	XAFSRTAMPRT4	Mixed	FB	EN+FR+DE	EN+FR+DE	0.2415	0.5414	0.4664	0.3165
12	xrce	ADDF	TXT	FB	EN+FR+DE	EN+FR+DE	0.2361	0.4871	0.4393	0.3077
13	unt	untaTxEn	TXT	NOFB	EN	EN	0.2251	0.4314	0.3871	0.3025
14	telecom	tefwm	TXT	QE	EN+FR+DE	EN	0.2227	0.4829	0.4407	0.2953
15	unt	untaTxFr	TXT	NOFB	FR	FR	0.2200	0.4229	0.3986	0.2855

Table 5: Results for the top 15 runs.

Rank	**Participant**	**Run**	**Modality**	**FB/QE**	**AL**	**TL**	**MAP**	**P@10**	**P@20**	**R-prec.**
1	xrce	XAFSQTAMP	MIXED	FB	EN+FR+DE	EN+FR+DE	0.2765	0.5814	0.5193	0.3465
13	unt	untaTxEn	TXT	NOFB	EN	EN	0.2251	0.4314	0.3871	0.3025
14	telecom	tefwm	TXT	QE	EN+FR+DE	EN	0.2227	0.4829	0.4407	0.2953
16	i2rcviu	MONOLINGUAL	TXT	NOFB	EN	EN+FR+DE	0.2126	0.4486	0.4143	0.2832
20	dcu	dcuRunOkapi	TXT	QE	EN	EN	0.2039	0.4271	0.3907	0.2832
24	cheshire	BTEA	TXT	FB	EN+FR+DE	EN+FR+DE	0.2014	0.4600	0.4036	0.2739
26	duth	D20MM	MIXED	NOFB	EN+FR+DE	EN+FR+DE	0.1998	0.5200	0.4836	0.2820
34	uned	UUEYANTT	TXT	NOFB	EN+FR+DE	EN+FR+DE	0.1927	0.3914	0.3564	0.2663
48	daedalus	DWCT	TXT	QE	EN+FR+DE	EN+FR+DE	0.1820	0.4471	0.4029	0.2662
50	sztaki	bat	MIXED	QE	EN	EN	0.1794	0.4857	0.4329	0.2318
66	nus	nustextonly	TXT	NOFB	EN	EN	0.1581	0.3529	0.3264	0.2386
100	rgu	combine	MIXED	NOFB	EN	EN	0.0617	0.2271	0.2129	0.1221
110	uaic	dfiush	MIXED	QE	EN+FR+DE	EN+FR+DE	0.0423	0.1543	0.1529	0.0744

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

[5] `http://www.imageclef.org/2010/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. Because there was only one visual only run among the top 90%, it was also discarded. Table 6 shows the average performance and standard deviation with respect to each modality. On average, the textual runs have a slightly better performance than multimodal ones with respect to all examined evaluation metrics (MAP, Precision at 20, and precision after R (= number of relevant) documents retrieved).

Table 6: Results per modality over all topics.

Modality	MAP		P@20		R-prec.	
	Mean	SD	Mean	SD	Mean	SD
All top 90% runs (112 runs)	0.1543	0.0641	0.3541	0.1	0.2213	0.0756
TXTIMG in top 90% runs (67 runs)	0.1504	0.0719	0.3519	0.1156	0.2148	0.0843
TXT in top 90% runs (45 runs)	0.1602	0.0505	0.3575	0.0728	0.2312	0.0598

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 up to 10 topics per class (i.e., easy, medium, hard, and very hard), together with the total number of topics per class. Out of 70 topics, 53 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. 21 topics were very hard to solve, with four of them ("woman in red dress", "building site", "horseman", "people laughing") having a $MAP < 0.05$ and being considered as unsolvable. The topic pool includes only four easy topics ("satellite image", "portrait of Jintao Hu", "ferrari red", "postage stamp"). A large number of the topics included in the easy and medium classes include a reference to a named entity ("Jintao Hu", "Ferrari", "ISS" or "Shiva") and, consequently, are easily retrieved with simple textual approaches. As for very hard topics, they often contain general terms ("woman", "people", "house" or "airplane"), 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 - the total number of topics per class is given in the table header. Up to 10 topics, the hardest ones in each class, are shown in the table.

easy (4 topics)	medium (13 topics)	hard (32 topics)	very hard (21 topics)
52 satellite image	**23** emoticon smiley	**69** red tomato	**64** woman in red dress
36 portrait of Jintao Hu	**9** flying hot air balloon	**45** snowy winter landscape	**55** building site
40 ferrari red	**49** surfing on waves	**22** sharks underwater	**10** horseman
63 postage stamp	**4** Christmas tree	**42** Paintings related to cubism	**32** people laughing
	12 DNA helix	**11** landline telephone	**67** white house with garden
	18 stars and galaxies	**27** active volcano with ash cloud	**48** civil airplane
	53 ISS international space station	**68** historic castle	**16** spider with cobweb
	25 Shiva painting or sculpture	**29** desert scenery	**31** yellow buses
	41 polar bear	**65** sea sunset or sunrise	**60** notes on music sheet
	24 Rorschach black and white	**14** videogames screenshot	**19** lochs in Scottland

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 (TXTIMG). Figure 2 shows the average performance on each topic for all, text-only and text-visual runs. The text-based runs outperform the text-visual ones in 37 out of the 70, are outperformed by mixed runs in 31 cases and have the same performances in 2 cases. This indicates that less than half of the topics benefit from a multi modal 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, 26 of the topics can be characterized as textual and 24 as visual. The remaining 20 topics, where no clear improvements are observed, are considered to be neutral.

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. There are small differences between the average number of words and example images for textual, neutral and visual topics. An important difference is observed for the number of relevant documents/topic, with a significantly higher number of such documents for visual topics compared to textual topics. This distribution of the number of relevant images indicates that a larger number of positive examples per query are needed for the visual features to be effective. Interestingly, the average mean average precision is distributed inversely, with a significantly higher average score for textual queries compared to visual ones (0.219 vs. 0.131). The distribution of the textual, visual and neutral topics over the classes expressing their difficulty shows that the vi-

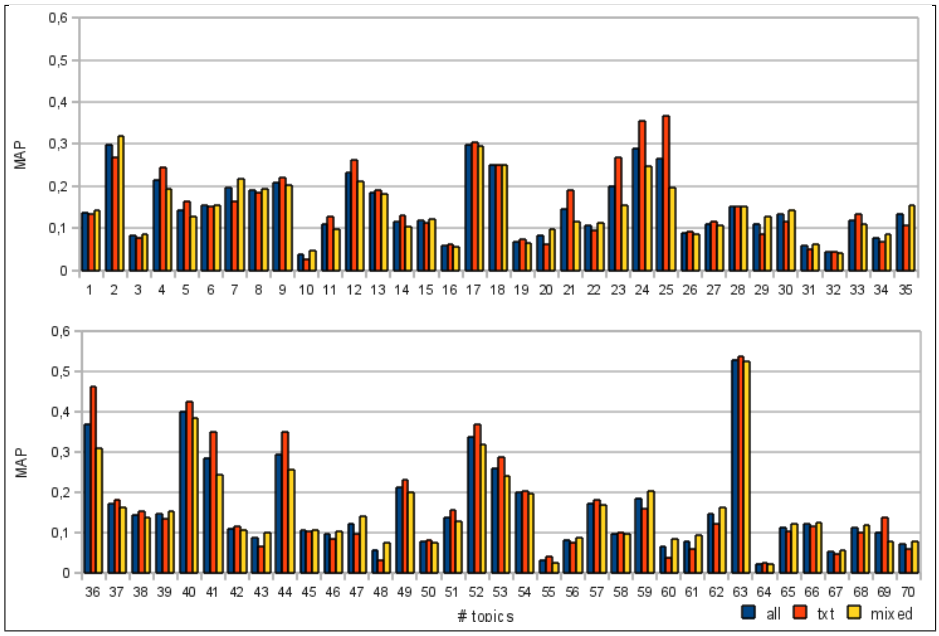

Fig. 2: Average topic performance over all, text-only, and mixed runs.

sual and neutral topics are more likely to fall into the hard/very hard class than the textual ones.

A closer look at the topics themselves indicates that textual ones often include a named entity ("ISS", "Ferrari", "Christmas", "Jintao Hu" etc.). Visual cues, which could be useful for topics that have a well defined semantic interpretation or a coherent visual aspect, do not help in these cases.

6.4 Effect of Query Expansion and Relevance Feedback

Finally, we analyze the effect of the application of query expansion (QE) and relevance feedback (FB) techniques. 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 for 34 topics and relevance feedback for 15. As with visual topics, query expansion seems to be useful for queries which have a lot of associated relevant documents and for queries that are either hard or very hard.

7 Conclusions

For the first time this year, a multimodal approach performed best in the Wikipedia Retrieval task. It is encouraging to see more than half of the submitted runs

Table 8: Best performing topics for textual and text-visual runs relative to the average over all runs.

	textual (21 topics)	**visual (36 topics)**	**neutral (13 topics)**
Topics	**9** flying hot air balloon	**1** fractals	**7** lightning in the sky
	69 red tomato	**10** horseman	**62** Shopping in a market
	53 ISS international space station	**13** trains and locomotives	**61** earth from space
	52 satellite image	**15** cyclist	**60** notes on music sheet
	51 aerial pictures of landscapes	**16** spider with cobweb	**59** cities at night
	5 Oktoberfest beer tent	**17** beach volleyball	**48** civil airplane
	49 surfing on waves	**18** stars and galaxies	**47** soccer stadium
	44 saturn	**19** lochs in Scottland	**43** skyscraper in daylight
	41 polar bear	**22** sharks underwater	**35** race car
	40 ferrari red	**26** brain scan	**30** harbour
	4 Christmas tree	**27** active volcano with ash cloud	**29** desert scenery
	37 close up of bottles	**28** palm trees	**20** mountains with sky
	36 portrait of Jintao Hu	**3** basketball game close up	**2** cockpit of an airplane
	33 close up of antenna	**31** yellow buses	
	25 Shiva painting or sculpture	**32** people laughing	
#images/topic	1.62	1.72	1.69
#words/topic	2.86	2.69	2.84
#reldocs	130.8	272.9	391.3
MAP	0.219	0.125	0.131
easy	3	1	0
medium	10	2	1
hard	8	17	7
very hard	0	16	5

were multimodal. This is possibly a consequence of the fact that visual descriptors were provided with the collection. A novelty this year was that participants were able to submit multilingual runs. Although a majority of runs focused either on a combination of topic languages or on English queries only, several groups submitted runs for German and French queries only.

8 Acknowledgements

Adrian Popescu was supported by the French ANR (Agence Nationale de la Recherche) vie the Georama project (ANR-08-CORD-009). Theodora Tsikrika was supported by the European Union via the European Commission project VITALAS (contract no. 045389). Jana Kludas was funded by the Swiss National Fund (SNF). The authors would also like to thank all the groups participating in the relevance assessment process.

The authors would like to thank Hervé le Borgne and Pierre-Alain Moëllic (CEA LIST) for providing the visual features for the collection.

References

1. Avi Arampatzis and Savvas A. Chatzichristofis and Konstantinos Zagoris Multimedia Search with Noisy Modalities: Fusion and Multistage Retrieval In *Working notes of the ImageCLEF 2010 Lab*, Padua, Italy, 2010.

Table 9: Best performing topics for query expansion (QE) and feedback (FB) runs relative to the average over all runs. Only the top 15 topics that benefit the from query expansion are presented here.

	QE (34 topics)	FB (15 topics)
Topics	**9** flying hot air balloon	**8** tennis player on court
	70 close up of trees	**69** red tomato
	7 lightning in the sky	**51** aerial pictures of landscapes
	68 historic castle	**5** Oktoberfest beer tent
	66 bridges in daylight	**44** saturn
	63 postage stamp	**43** skyscraper in daylight
	6 solar panels	**4** Christmas tree
	58 red fruits	**36** portrait of Jintao Hu
	57 road street signs	**33** close up of antenna
	56 musician on stage	**25** Shiva painting or sculpture
	54 launching space shuttle	**24** Rorschach black and white
	52 satellite image	**22** sharks underwater
	51 aerial pictures of landscapes	**21** Chernobyl disaster ruins
	50 portraits of people	**18** stars and galaxies
	46 sailboat	**11** landline telephone
#images/topic	2.62	3
#words/topic	1.79	1.8
#reldocs	353.3	144.2
avg. MAP	0.16	0.188
easy	2	1
medium	5	5
hard	18	8
very hard	9	1

2. J. Benavent and X. Benavent and E. de Ves and R. Granados and Ana Garca-Serrano Experiences at ImageCLEF 2010 using CBIR and TBIR mixing information approaches In *Working notes of the ImageCLEF 2010 Lab*, Padua, Italy, 2010.
3. Ya-Chun Cheng and Shu-Yuan Chen. Image classification using color, texture and regions, *Image and Vision Computing, Volume 21, 759-776*, 2003.
4. Stéphane Clinchant and Gabriela Csurka and Julien Ah-Pine and Guillaume Jacquet and Florent Perronnin and Jorge Sánchez and Keyvan Minoukadeh XRCE's Participation in Wikipedia Retrieval, Medical Image Modality Classifiction and Ad-hoc Retrieval Tasks of ImageCLEF 2010 In *Working notes of the ImageCLEF 2010 Lab*, Padua, Italy, 2010.
5. Bálint Daróczy and István Petrás and András A. Benczúr SZTAKI @ ImageCLEF 2010 In *Working notes of the ImageCLEF 2010 Lab*, Padua, Italy, 2010.
6. Bertrand Delezoide, Hervé Le Borgne, Romaric Besançon, Gaël de Chalendar, Olivier Ferret, Faiza Gara, Patrick Hède, Meriama Laib, Olivier Mesnard, Pierre-Alain Moëllic and Nasredine Semmar MM: modular architecture for multimedia information retrieval In *8^{th} International Workshop on Content-Based Multimedia Indexing (CBMI 2010, demo session)*, Grenoble, France, 2010.
7. Sara Lana-Serrano, Julio Villena-Román, José Carlos González-Cristóbal DAEDALUS at ImageCLEF Wikipedia Retrieval 2010: Expanding with Semantic Information from Context In *Working notes of the ImageCLEF 2010 Lab*, Padua, Italy, 2010.
8. Ray R. Larson Blind Relevance Feedback for the ImageCLEF Wikipedia Retrieval Task In *Working notes of the ImageCLEF 2010 Lab*, Padua, Italy, 2010.
9. Jinming Min and Johannes Leveling and Gareth Jones Document Expansion for Text-based Image Retrieval at WikipediaMM 2010 In *Working notes of the ImageCLEF 2010 Lab*, Padua, Italy, 2010.

10. Carol Peters and Theodora Tsikrika and Henning Müller and Jayashree Kalpathy-Cramer, Gareth J.F.Jones, Julio Gonzalo, and Barbara Caputo, editors. *Multilingual Information Access Evaluation Vol. II Multimedia Experiments: Proceedings of the 10th Workshop of the Cross-Language Evaluation Forum (CLEF 2009)*, Lecture Notes in Computer Science. Springer, 2010.
11. Jun Wang and Dawei Song and and Leszek Kaliciak RGU at ImageCLEF2010Wikipedia Retrieval Task In *Working notes of the ImageCLEF 2010 Lab*, Padua, Italy, 2010.
12. Adrian Popescu Télécom Bretagne at ImageCLEF WikipediaMM 2010 In *Working notes of the ImageCLEF 2010 Lab*, Padua, Italy, 2010.
13. Miguel E. Ruiz, Jianping Chen, Karthikeyan Pusapathy, Pok Chin and Ryan Knudson UNT at ImageCLEF 2010: CLIR for Wikipedia Images In *Working notes of the ImageCLEF 2010 Lab*, Padua, Italy, 2010.
14. Josef Sivic and Andrew Zisserman. Video Google: Efficient Visual Search for Videos, Towards Category-Level Object Recognition, *LNCS 4170*, pages 127144 , 2006.
15. Cees G. M. Snoek, Marcel Worring, Jan C. van Gemert, Jan-Mark Geusebroek, and Arnold W. M. Smeulders. The challenge problem for automated detection of 101 semantic concepts in multimedia. In *Proceedings of the 14th annual ACM international conference on Multimedia*, pages 421–430, New York, NY, USA, 2006. ACM Press.
16. Renato O. Stehling, Mario A. Nascimento and Alexandre A. Falcão A compact and efficient image retrieval approach based on border/interior pixel classification In *Proceedings of the 11th International Conference on Information and Knowledge Management*, pages 102–109, McLean, VA, USA, 2002.
17. Srinivasarao Vundavalli. IIIT-H at ImageCLEF Wikipedia MM 2009. In *CLEF 2009 working notes*, 2009.
18. Kong-Wah Wan and Yan-Tao Zheng, Sujoy Roy I2R at ImageCLEF Wikipedi Retrieval 2010 In *Working notes of the ImageCLEF 2010 Lab*, Padua, Italy, 2010.
19. Thijs Westerveld and Roelof van Zwol. The INEX 2006 multimedia track. In Norbert Fuhr, Mounia Lalmas, and Andrew Trotman, editors, *Advances in XML Information Retrieval: 5th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2006, Revised Selected Papers*, volume 4518, pages 331–344. Springer, 2007.