Chapter 9
The Wikipedia Image Retrieval Task

Theodora Tsikrika and Jana Kludas

Abstract The Wikipedia image retrieval task at ImageCLEF provides a test–bed for the system–oriented evaluation of visual information retrieval from a collection of Wikipedia images. The aim is to investigate the effectiveness of retrieval approaches that exploit textual and visual evidence in the context of a large and heterogeneous collection of images that are searched for by users with diverse information needs. This chapter presents an overview of the available test collections, summarises the retrieval approaches employed by the groups that participated in the task during the 2008 and 2009 ImageCLEF campaigns, provides an analysis of the main evaluation results, identifies best practices for effective retrieval, and discusses open issues.

9.1 Introduction

The Wikipedia image retrieval task, also referred to as the WikipediaMM task, is an ad hoc image retrieval task whereby retrieval systems are given access to a collection of images to be searched but cannot anticipate the particular topics that will be investigated. The image collection consists of freely distributable Wikipedia[1] images annotated with user–generated textual descriptions of varying quality and length. Given a user's multimedia information need expressed both as a textual query and also through visual cues in the form of one or more sample images or visual concepts, the aim is to find as many relevant images as possible. Retrieval approaches should exploit the available textual and visual evidence, either in isolation or in combination, in order to achieve the best possible ranking for the user.

Theodora Tsikrika
Centrum Wiskunde & Informatica, Science Park 123, 1098 XG Amsterdam, The Netherlands
e-mail: theodora.tsikrika@acm.org

Jana Kludas
CUI, University of Geneva, Switzerland e-mail: Jana.Kludas@unige.ch

[1] http://www.wikipedia.org/

H. Müller et al. (eds.), *ImageCLEF*, The Information Retrieval Series 32,
DOI 10.1007/978-3-642-15181-1_9,

The task was set up in 2006 as part of the activities of the INEX Multimedia track (Westerveld and van Zwol, 2007), where it was referred to as the MMimages task. In 2008, the task moved to ImageCLEF, which not only forms a more natural environment for hosting this type of benchmark but also attracts more participants from the content–based image retrieval community. The overall goal of the task is to promote progress in large scale, multi–modal image retrieval via the provision of appropriate test collections that can be used to reliably benchmark the performance of different retrieval approaches using a metrics–based evaluation.

This chapter presents an overview of the Wikipedia image retrieval task in the ImageCLEF 2008 and 2009 evaluation campaigns (Tsikrika and Kludas, 2009, 2010). Section 9.2 presents the evaluation objectives of this task and describes the task's resources, i.e. the Wikipedia image collection and additional resources, the topics, and the relevance assessments. Section 9.3 lists the research groups that participated in these two years of the task under ImageCLEF, outlines the approaches they employed, and presents the results of the evaluation. Section 9.4 examines the results achieved by specific approaches in more detail so as to identify best practices and discuss open issues. Section 9.5 concludes this chapter, provides information on how to access the available resources, and discusses the future of the task.

9.2 Task Overview

9.2.1 Evaluation Objectives

The Wikipedia image retrieval task during the ImageCLEF 2008 and 2009 campaigns aimed to provide appropriate test collections for fostering research towards the following objectives:

Firstly, the task aimed to investigate how well image retrieval approaches, particularly those that exploit visual features, could deal with larger scale image collections. To this end, the goal was to provide a collection of more than 150,000 images; such a collection would be, for instance, much larger than the IAPR TC–12 image collection (Grubinger et al, 2006) that consists of 20,000 photographs and that was, at the time, employed in the ImageCLEF 2008 photo retrieval task (Arni et al, 2009).

Secondly, it aimed to examine how well image retrieval approaches could deal with a collection that contains highly heterogeneous items both in terms of their textual descriptions and their visual content. The textual metadata accompanying the Wikipedia images are user–generated, and thus outside any editorial control and correspond to noisy and unstructured textual descriptions of varying quality and length. Similarly, Wikipedia images cover highly diverse topics and since they are also contributed by Wikipedia users, their quality cannot be guaranteed. Such characteristics pose challenges for both text–based and visual–based retrieval approaches.

Finally, the main aim was to study the effectiveness of retrieval approaches that combine textual and visual evidence in order to satisfy a user's multimedia information need. Textual approaches had proven hard to beat in well–annotated image collections. However, such collections are not the norm in realistic settings, particularly in the Web environment. Therefore, there was a need to develop multi–modal approaches able to leverage all available evidence.

9.2.2 Wikipedia Image Collection

The collection of Wikipedia images used in the Wikipedia image retrieval task during the 2008 and 2009 ImageCLEF campaigns is a cleaned–up version of the image collection created in 2006 in the context of the activities of the INEX Multimedia track, where it was employed for the MMimages task in 2006 (Westerveld and van Zwol, 2007) and 2007 (Tsikrika and Westerveld, 2008). Due to its origins, the collection is referred to as the (INEX MM) Wikipedia image collection.

This image collection was created out of the more than 300,000 images contained within the 659,388 English Wikipedia articles that were downloaded and converted to XML (Denoyer and Gallinari, 2007) so as to form the structured document collection used for the ad hoc and other tasks at INEX 2006 (Malik et al, 2007). The user-generated metadata accompanying these Wikipedia images, usually a brief caption or description of the image, the Wikipedia user who uploaded the image, and the copyright information, were then downloaded and also converted to XML. Due to copyright issues or parsing problems with the downloaded metadata, some images had to be removed leaving a collection of approximately 170,000 images that was used in the INEX Multimedia tracks of 2006 (Westerveld and van Zwol, 2007) and 2007 (Tsikrika and Westerveld, 2008). Once the task became part of ImageCLEF in 2008, the collection was further cleaned up with the aim of keeping only JPEG and PNG images, leading to a collection of 151,519 diverse images with highly heterogeneous and noisy textual descriptions of varying length.

9.2.3 Additional Resources

To encourage participants to investigate multi-modal approaches that combine textual and visual evidence, particularly research groups with expertise only in the field of textual Information Retrieval, a number of additional resources were also provided.

In 2008, the following resources, computed during the INEX 2006 Multimedia track, were made available to support researchers who wished to exploit visual evidence without performing image analysis:

Image classification scores: For each image in the collection, the classification scores for the 101 MediaMill concepts were provided by the University of Am-

sterdam (Snoek et al, 2006). Their classifiers had been trained on manually annotated TREC Video Retrieval Evaluation (TRECVID) video data for concepts selected for the broadcast news domain.

Visual features: For each image in the collection, the set of the 120D feature vectors that had been used to derive the above image classification scores (van Gemert et al, 2006) were also made available. Participants could use these feature vectors to custom–build a content–based image retrieval system, without having to pre–process the image collection.

In 2009, the following resource was added:

Image similarity matrix: The similarity matrix for the images in the collection was constructed by the IMEDIA group at INRIA. For each image in the collection, this matrix contains the list of the top $K = 1,000$ most similar images in the collection together with their similarity scores. The same was given for each image used as a query example in the topics. The similarity scores are based on the distance between images; therefore, the lower the score, the more similar the images. Further details on the features and distance metric used can be found in Ferecatu (2005).

9.2.4 Topics

Topics are descriptions of multimedia information needs, with each topic containing textual and visual cues that can be used as evidence of the relevance of the images that should be retrieved. A number of factors have to be taken into consideration when creating topics for a test collection since such topics should reflect the real needs of operational retrieval systems, represent the types of services such systems might provide, be diverse, and differ in their coverage.

In 2008, the Wikipedia image retrieval task adopted the topic creation process introduced in INEX, whereby all participating groups were required to submit candidate topics. The participants were provided with topic development guidelines (Kludas and Tsikrika, 2008) which were based on guidelines created earlier in the context of INEX tasks (Larsen and Trotman, 2006). The participating groups submitted 70 topics altogether, which, together with 35 topics previously used in the INEX 2006 and 2007 Multimedia track, formed a pool of 105 candidate topics. Out of these, the task organisers selected a set of 75 topics. In 2009, participation in the topic development process was not mandatory, so only two of the participating groups submitted a total of 11 candidate topics. The rest of the candidate topics were created by the organisers with the help of the log of an image search engine. After a selection process performed by the organisers, a final list of 45 topics was created.

The topics consist of the following parts:

<**title**> query by keywords,
<**image**> query by image examples (one or more) — *optional in 2008*,
<**concept**> query by visual concepts (one or more) — *only in 2008 and optional*,

<**narrative**> definitive description of relevance and irrelevance.

The topic's <title> simulates a user who does not have (or does not want to use) example images or other visual cues. The query expressed in the <title> is therefore a text–only query. Upon discovering that a text–only query does not produce many relevant results, a user might decide to add visual cues and formulate a multimedia query. The topic's <image> provides visual cues that correspond to example images taken from outside or inside the (INEX MM) Wikipedia image collection and can be of any common format. In 2008, it was optional for topics to contain such image examples, whereas in 2009, each of the topics had at least one, and in many cases several, example images that could help describe the visual diversity of the topic. In 2008, additional visual cues were provided in the <concept> field that contained one or more of the 101 MediaMill concepts for which classification scores were provided.

These textual and visual evidences of relevance can be used in any combination by the retrieval systems; it is up to them how to use, combine or ignore this information. The relevance of a result does not directly depend on these constraints, but is decided by manual assessments based on the <narrative>. This field is not provided to the participants, but only to the assessors, and contains a clear and precise description of the information need in order to unambiguously determine whether or not a given image fulfils the given information need. The <narrative> 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. To aid this, the <narrative> should explain not only what information is being sought, but also the context and motivation of the information need, i.e. why the information is being sought and what work–task it might help to solve.

Table 9.1 lists some statistics for the topics that were used during these two years of the task. The titles of these topics can be found in the overview papers of the task (Tsikrika and Kludas, 2009, 2010). The topics range from simple and thus relatively easy (e.g. 'bikes') to semantic and hence highly difficult (e.g. 'aerial photos of non–artificial landscapes'), 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; they were created so as to be challenging for current state–of–the–art retrieval algorithms. As mentioned above, in 2008, not all topics contained visual cues since the aim was to represent scenarios where users expressing their multimedia information needs do not necessarily employ visual evidence.

9.2.5 Relevance Assessments

In the Wikipedia image retrieval task, each image was assessed either as being relevant or as being non relevant, i.e. binary relevance was assumed. The retrieved images contained in the runs submitted by the participants were pooled together using a pool depth of 100 in 2008, which resulted in pools that ranged from 753 to

Table 9.1: Statistics for the topics in the ImageCLEF 2008 and 2009 Wikipedia image retrieval.

	2008	2009
Number of topics	75	45
Average number of terms in title	2.64	2.7
Average number of images per topic	0.61	1.9
Number of topics with image(s)	43	45
Number of topics with concept(s)	45	–
Number of topics with both image(s) and concept(s)	28	–
Number of topics with title only	15	–

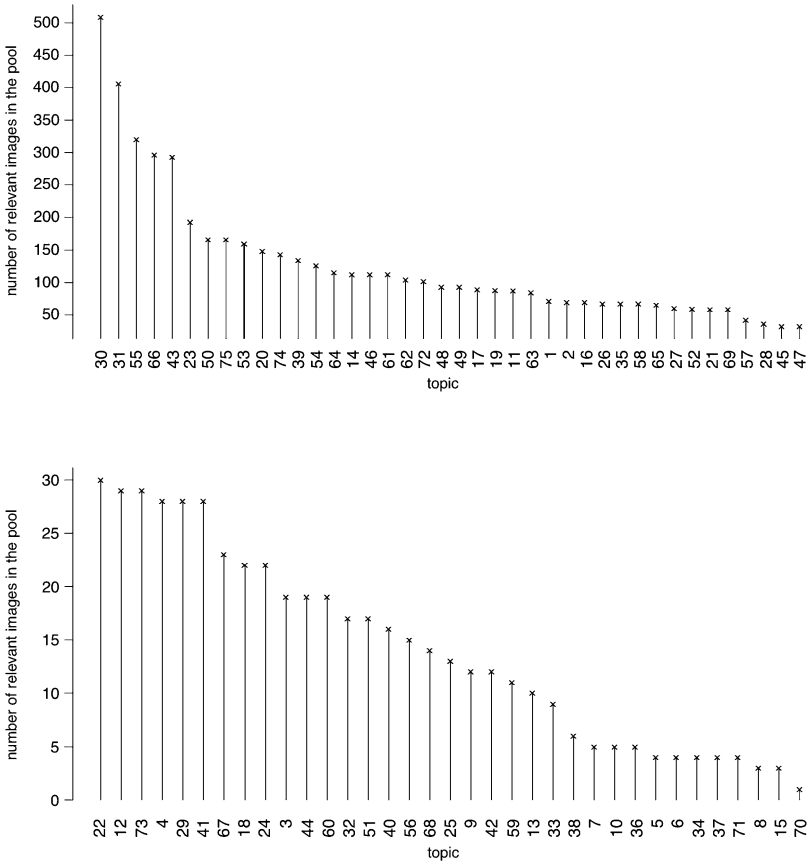


Fig. 9.1: Number of relevant images for each of the 2008 topics; topics are sorted in decreasing order of the number of their relevant images.

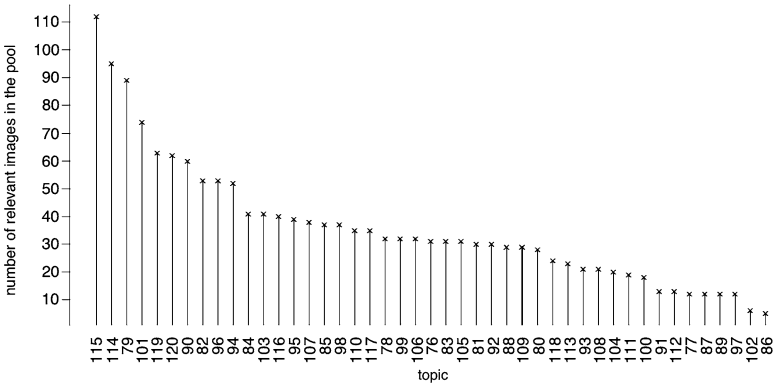


Fig. 9.2: Number of relevant images for each of the 2009 topics; topics are sorted in decreasing order of the number of their relevant images.

1,850 images with a mean and median both around 1,290, and a pool depth of 50 in 2009, which resulted in pools that ranged from 299 to 802 images with a mean and median both around 545. The evaluation was performed by the participants of the task within a period of four weeks after the submission of runs: 13 groups participated in 2008 and seven groups in 2009. The assessors used a Web–based relevance assessment system that had been previously employed in the INEX Multimedia and TREC Enterprise tracks (see Chapter 4 in this volume for further information on this system). In 2008, given that most topics were created by the participants, who were also employed as assessors, an effort was made so as to ensure that most of the topics were assigned to their creators. This was achieved in 76% of the assignments of the topics that were created that year.

Figures 9.1 and 9.2 depict the distribution of relevant images in the judged pools for each of the topics in 2008 and 2009, respectively. The variability in the number of relevant images across topics is evident, with most topics though having less than 100 relevant images. The mean number of relevant images per topic is 74.6 for 2008 and 36 for 2009, while the median is 36 and 31, respectively. Over all 120 topics, the mean number of relevant images per topic is 60.1 and the median 32.

9.3 Evaluation

9.3.1 Participants

Compared to its previous incarnation in the context of the INEX Multimedia track, the Wikipedia image retrieval task attracted more interest once it moved under the

Table 9.2: Groups that participated in the Wikipedia image retrieval task during the 2008 and 2009 ImageCLEF campaigns. Each entry lists the group ID, the academic or research institute hosting the group, the country where it is located, and the number of runs the group submitted in each of the campaigns.

Group ID	Institution	Country	2008	2009
cea	CEA-LIST	France	2	12
chemnitz	Chemnitz University of Technology	Germany	4	–
cwi	Centrum Wiskunde & Informatica	Netherlands	2	–
dcu	Dublin City University	Ireland	–	5
deuceng	Dokuz Eylul University	Turkey	–	6
iiit-h	International Institute of Information Technology, Hyderabad	India	–	1
imperial	Imperial College	UK	6	–
irit	Institut de Recherche en Informatique de Toulouse	France	4	–
lahc	Université Jean Monnet, Saint–Étienne	France	6	13
sinai	University of Jaen	Spain	–	4
sztaki	Hungarian Academy of Science	Hungary	8	7
ualicante	University of Alicante	Spain	24	9
unige	Université de Genève	Switzerland	2	–
upeking	Peking University	China	7	–
upmc/lip6	LIP6, Université Pierre et Marie Curie	France	7	–
utoulon	Université Sud Toulon–Var	France	5	–
		Total runs	77	57

auspices of ImageCLEF. The number of groups that participated by submitting runs was 12 in 2008 and eight in 2009, four of which were returning participants. Table 9.2 lists the participating groups along with the number of runs they submitted for the official evaluation; a total of 77 runs were submitted in 2008, while 57 runs were submitted in 2009. The overwhelming majority of participants are based in Europe, with the exception of only two groups, one from China and one from India.

9.3.2 Approaches

The approaches employed by the participants have been quite diverse. Both textual and visual features have been considered, either in isolation or in combination. Query and document expansion techniques that exploit semantic knowledge bases have been widely applied, as well as query expansion approaches that rely on blind relevance feedback. A short description of the participants' approaches is provided next. Each group is represented by its ID, followed by the year(s) in which the group participated in the task, and the publication(s) where the employed approaches are described in more detail. The groups are listed in alphabetical order of their ID.

cea (2008, 2009) (Popescu et al, 2009; Myoupo et al, 2010) In 2008, they employed Wikipedia and WordNet[2] as knowledge bases for automatically identifying and ranking concepts considered to be semantically related to those in the textual part of the query topics. These concepts were used for expanding the query, which was then submitted against the index of the images' textual descriptions, so as to generate a text–based ranking. In their visual analysis, the images in the collection were classified with respect to several visual concepts using Support Vector Machine (SVM)–based classifiers that exploited colour histogram and texture Local–Edge Pattern (LEP) visual features. Textual concepts in the queries triggered the use of visual concepts (e.g., persons' names triggered the use of the face detector) and the images' classification scores for these concepts were used for re–ranking the text–based results. In 2009, they refined the textual query expansion process by using knowledge extracted only from Wikipedia, whereas for the visual re–ranking they introduced a k–Nearest Neighbour (k–NN) based method. This method builds a visual model of the query using the top–ranked images retrieved by Google[3] and Yahoo![4] for that query and re–ranks the images in the text–based results based on their visual similarity to the query model.

chemnitz (2008) (Wilhelm et al, 2008) They employed their Xtrieval framework, which is based on Lucene[5] and PostgreSQL[6], and considered both textual and visual features, as well as the provided resources (image classification scores and low–level visual features). The text–based retrieval scores were combined with the visual similarity scores and further combined with the concept–based image classification scores. A thesaurus–based query expansion approach was also investigated.

cwi (2008) (Tsikrika et al, 2008) They employed PF/Tijah[7], an XML retrieval framework for investigating a language modelling approach based on purely textual evidence. A length prior was also incorporated so as to bias retrieval towards images with longer descriptions than the ones retrieved by the language model.

deuceng (2009) (Kilinc and Alpkocak, 2009) They applied a two–step approach: 1) text–based retrieval using expanded image descriptions and queries, and 2) re–ranking based on Boolean retrieval and text–based clustering. Terms and term phrases in both image descriptions and queries were expanded using WordNet, through the application of word sense disambiguation and WordNet similarity functions. The text–based results generated in this first step were then re–ranked in a Boolean manner by boosting the scores of the images that contained in their descriptions all the query terms in the exact same order as the query. The vectors of textual features of the results generated in the first step together with the vector of the expanded query were then clustered using the cover coefficient–

[2] http://wordnet.princeton.edu/

[3] http://images.google.com/

[4] http://images.search.yahoo.com/

[5] http://lucene.apache.org/

[6] http://www.postgresql.org/

[7] http://dbappl.cs.utwente.nl/pftijah/

based clustering methodology (C^3M). This allowed the calculation of similarity scores between the query vector and the vectors of the retrieved images. The final score was computed as a weighted sum of the Boolean re–ranking and the C^3M re–ranking scores. For further details, see Chapter 14 in this volume.

dcu (2009) (Min et al, 2010) They focused their experimentations on the expansion of the images' textual descriptions and of the textual part of the topics, using the Wikipedia abstracts' collection DBpedia[8] and blind relevance feedback. When DBpedia was employed, the terms from its top–ranked documents retrieved in response to the image description (or textual query) were sorted by their frequency and the top–ranked were selected to expand the images' (or queries') text. The term re–weighting was performed using Rocchio's formula. Query expansion was also performed using blind relevance feedback and BM25 term re–weighting. Lemur[9] was employed as the underlying retrieval framework.

iiit-h (2009) (Vundavalli, 2009) They employed a simple text–based approach that first used Boolean retrieval so as to narrow down the collection to the images accompanied by descriptions that contained all query terms and then ranked these images by applying the vector space model using a *tf.idf* weighting scheme.

imperial (2008) (Overell et al, 2008) They examined textual features, visual features, and their combination. Their text-based approach also took into account evidence derived from a geographic co-occurrence model mined from Wikipedia which aimed at disambiguating geographic references in a context-independent or a context-dependent manner. Their visual-based approach employed Gabor texture features and the City Block distance as a similarity measure. Text-based and visual-based results were combined using a convex combination of ranks. The results of this combination were further merged with results generated from using the top-ranked text-based results as blind relevance feedback in their visual retrieval approach.

irit (2008) (Torjmen et al, 2009) They explored the use of image names as evidence in text-based image retrieval. They first used them in isolation by computing a similarity score between the query and the name of the images in the collection using the vector space model. Then they used them in combination with textual evidence either by linearly combining the ranking of their text-based approach implemented in their XFIRM retrieval system with the ranking produced by the name-based technique or by applying a text-based approach that boosts the weights of terms that also occur in the image name.

lahc (2008, 2009) (Moulin et al, 2009, 2010) In 2008, they used a vector space model to compute similarities between vectors of both textual and visual terms. The textual terms corresponded to textual words and their weights were computed using BM25. The visual terms were obtained through a bag of words approach and corresponded to six–dimensional vectors of clusters of local colour features extracted from the images and quantized by k–means. Both manual

[8] http://dbpedia.org/

[9] http://www.lemurproject.org/

and blind relevance feedback were applied to a text–based run so as to expand the query with visual terms. In 2009, their document model was simplified so as to consider textual and visual terms separately and their approach was extended as follows. Additional textual information was extracted from the original Wikipedia articles that contained the images. Several local colour and texture features, including Scale Invariant Feature Transform (SIFT) descriptors, were extracted. Finally, the text–image combination was now performed by linearly combining the text–based and visual–based rankings.

sinai (2009) (Díaz-Galiano et al, 2010) Their approach focused on the expansion of the images' textual descriptions and of the textual part of the topics using WordNet. All nouns and verbs in the image descriptions and text queries were expanded by adding all unique words from all of their WordNet synsets without applying any disambiguation. Lemur was employed as the underlying retrieval framework.

sztaki (2008, 2009) (Racz et al, 2008; Daróczy et al, 2009) In 2008, they used their own retrieval system developed by the Hungarian Academy of Sciences and experimented with a text–based approach that used BM25 and query expansion based on Local Context Analysis (LCA), and its linear combination with a segment–based visual approach. In 2009, they preprocessed the textual image descriptions in order to remove author and copyright information with the aim to reduce the noise in the index. Their text–based approach again used BM25, but query expansion was performed by employing an on–line thesaurus. Their visual runs employed image segmentation and SIFT descriptors. The text–based and visual–based rankings were linearly combined to produce the final score.

ualicante (2008, 2009) (Navarro et al, 2008, 2009) In 2008, they employed their textual passage–based IR–n retrieval system as their baseline approach which was enhanced 1) by a module that decomposed the (compound) image file names in camel case notation into single terms, and 2) by a module that performed geographical query expansion. They also investigated two different term selection strategies for query expansion: probabilistic relevance feedback and local context analysis. In 2009, they further extended their approach by also using the top–ranked images (and their textual descriptions) returned by a content–based visual retrieval system as input for the above term selection strategies performing text–based query expansion.

unige (2008) They employed only textual features and their approach was based on the preference ranking option of the SVM light library developed by Cornell University. One run also applied feature selection to the high dimensional textual feature vector, based on the features relevant to each query.

upeking (2008) (Zhou et al, 2009) They investigated the following approaches: 1) a text–based approach based on the vector space model with *tf.idf* term weights, also using query expansion where the expansion terms were automatically selected from a knowledge base that was (semi–)automatically constructed from Wikipedia, 2) a content–based visual approach, where they first trained 1 vs. all classifiers for all queries by using the training images obtained by Yahoo! image search and then treated the retrieval task as a visual concept detection in

Runs	2008	2009
Textual	35	26
Visual	5	2
Mixed	37	29
All	77	57

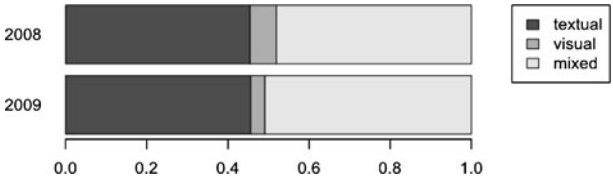


Fig. 9.3: Distribution of runs that employed textual, visual, or a combination of textual and visual low level features over the two years of the Wikipedia image retrieval task.

the given Wikipedia image set, and 3) a cross–media approach that combined the textual and visual rankings using the weighted sum of the retrieval scores.

umpc/lip6 (2008) (Fakeri-Tabrizi et al, 2008) They investigated text–based image retrieval by using a *tf.idf* approach, a language modelling framework, and their combination based on the ranks of retrieved images. They also experimented with the combination of textual and visual evidence by re–ranking the text–based results using visual similarity scores computed by either the Euclidean distance or a manifold–based technique, both on Hue/Saturation/Value (HSV) features.

utoulon (2008) (Zhao and Glotin, 2008) They applied the same techniques they used for the visual concept detection task at ImageCLEF 2008 (see Chapter 11 in this volume for details of that task) by relating each of the topics to one or more visual concepts from that task. These visual–based rankings were also fused with the results of a text–based approach.

All these different approaches can be classified with respect to whether they employ textual or visual low level features or a combination of both; in the latter case, an approach is characterised as mixed. Half of the groups that participated over the two years (eight out of the 16 groups) employed mixed approaches, whereas the other half relied only on textual features. Figure 9.3 shows the distribution of the submitted runs over the types of features they used. In both years of the task, mixed runs had a very slight edge over the textual runs.

The description of the runs submitted by the various groups also reveals that query expansion has been a very popular strategy as it has been applied by 11 of the 16 groups, either through the use of existing or purpose–built semantic knowledge bases (six out of the 11 groups), or through blind relevance feedback that takes into account textual or visual features (three out of the 11 groups), or as a combination of both these techniques (two out of the 11 groups). The application of query expansion aims to deal with the vocabulary mismatch problem, an issue which is particularly prominent in this test collection given both the short textual descriptions accompanying the images and the small number of images provided as query examples. A similar approach that has been applied by three out of the 16 groups with the aim to enrich the available textual descriptions of the Wikipedia images has been document

expansion with the use of semantic knowledge bases. Next, the results of the runs submitted by the participating groups over the two years of the task are presented.

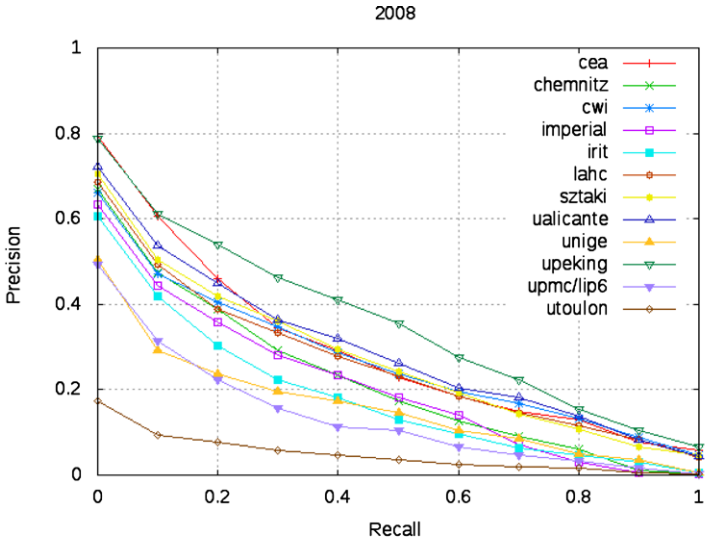
9.3.3 Results

The effectiveness of the submitted runs has been evaluated using the following measures: Mean Average Precision (MAP), P@10, P@20, and R–precision, i.e. precision when R (=number of relevant) documents are retrieved; see Chapter 5 in this volume for further details on these evaluation measures.

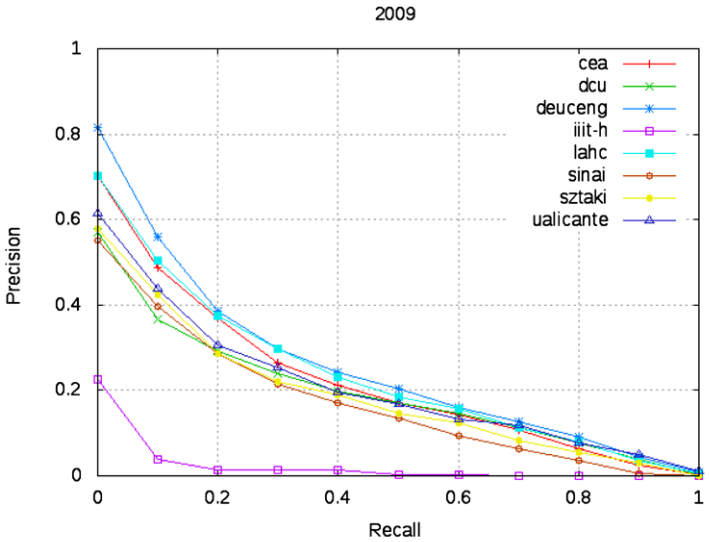
Figure 9.4 presents the best submitted run for each of the participating groups. Overall, the groups performed slightly better in 2008, an indication perhaps that the 2009 topics were more challenging for the participants. The best performing groups, *upeking* and *cea* in 2008, and *deuceng* in 2009, all employed query expansion, with the latter also performing document expansion, using semantic knowledge bases, such as WordNet and information extracted from Wikipedia. This indicates the usefulness of this approach in this particular setting. Furthermore, the best performing run both in 2008 and in 2009 relied only on textual evidence. This is better illustrated in Table 9.3 that presents a more complete overview of the submitted runs.

Table 9.3 shows for all runs, as well as for only the textual, visual, and mixed runs submitted in a year, the best, worst, median, and mean achieved performance for various evaluation measures. Both in 2008 and in 2009, the best values were achieved by runs that exploit only the available textual evidence. However, the differences between the best textual and the best mixed run for 2008 are not statistically significant for P@10 and P@20 ($p < 0.05$). Furthermore, the differences between the best textual and the best mixed run for 2009 are not statistically significant for all of the reported evaluation measures. On average, the median performance achieved by a mixed run in 2008 is slightly better than the median performance achieved by a textual run in terms of MAP and R–precision, while in 2009 the median values of all reported evaluation measures are higher for the mixed compared to the textual runs. On the other hand, the performance of the visual–only runs is comparatively low.

Given that a number of different evaluation measures were reported, a question that can be raised is whether there are any differences in these measures with respect to how they rank the submitted runs. To investigate this issue, the correlations among these measures were computed using the methodology described by Buckley and Voorhees (2005). For each evaluation measure, the runs are first ranked in order of decreasing performance with respect to that measure. The correlation between any two measures is then defined as the Kendall's τ correlation between the respective rankings. Table 9.4 presents the results of this analysis, where in addition to the evaluation measures previously reported, i.e. MAP, P@10, P@20, and R–precision, the total number of relevant images retrieved (abbreviated as 'Rel ret'), i.e. the sum of the number of relevant images retrieved across all topics for



(a) The best retrieval results per group for the 2008 Wikipedia image retrieval task.



(b) The best retrieval results per group for the 2009 Wikipedia image retrieval task.

Fig. 9.4: The best retrieval results per group.

a year, is also reported. The correlations between the MAP, P@10, P@20, and R–precision measures are all at least 0.67 showing that each pair of measures is corre-

Table 9.3: The best, worst, median and mean performance achieved by all, text only, visual only, and mixed only runs for MAP, P@10, P@20, and R–precision in the 2008 and the 2009 Wikipedia image retrieval tasks. The standard deviation of the performance achieved by the runs in each case is also listed.

		2008				2009			
		MAP	P@10	P@20	R-prec.	MAP	P@10	P@20	R-prec.
All runs		77 runs				57 runs			
	max	0.3444	0.4760	0.3993	0.3794	0.2397	0.4000	0.3189	0.2708
	min	0.0010	0.0027	0.0033	0.0049	0.0068	0.0244	0.0144	0.0130
	median	0.2033	0.3053	0.2560	0.2472	0.1699	0.2644	0.2267	0.2018
	mean	0.1756	0.2761	0.2230	0.2122	0.1578	0.2624	0.2153	0.1880
	stdev	0.0819	0.1169	0.0936	0.0920	0.0571	0.0861	0.0702	0.0631
Textual runs		35 runs				26 runs			
	max	0.3444	0.4760	0.3993	0.3794	0.2397	0.4000	0.3189	0.2708
	min	0.0399	0.0467	0.0673	0.0583	0.0186	0.0689	0.0389	0.0246
	median	0.2033	0.3107	0.2587	0.2472	0.1680	0.2600	0.2178	0.1987
	mean	0.1953	0.2972	0.2453	0.2356	0.1693	0.2717	0.2232	0.1992
	stdev	0.0662	0.0859	0.0690	0.0684	0.0452	0.0717	0.0574	0.0487
Visual runs		5 runs				2 runs			
	max	0.1928	0.4507	0.3227	0.2309	0.0079	0.0222	0.0222	0.0229
	min	0.0010	0.0027	0.0033	0.0049	0.0068	0.0144	0.0144	0.0130
	median	0.0037	0.0147	0.0120	0.0108	0.0074	0.0183	0.0183	0.0179
	mean	0.0781	0.1848	0.1336	0.0962	0.0074	0.0183	0.0183	0.0179
	stdev	0.1039	0.2415	0.1726	0.0122	0.0008	0.0055	0.0055	0.0070
Mixed runs		37 runs				29 runs			
	max	0.2735	0.4653	0.3840	0.3225	0.2178	0.3689	0.2867	0.2538
	min	0.0053	0.0040	0.0047	0.0049	0.0321	0.1044	0.0644	0.0423
	median	0.2083	0.3053	0.2547	0.2536	0.1801	0.2778	0.2389	0.2103
	mean	0.1701	0.2684	0.2139	0.2056	0.1578	0.2706	0.2218	0.1898
	stdev	0.0841	0.1172	0.0949	0.0967	0.0543	0.0776	0.0063	0.0605

Table 9.4: Kendall's τ correlations between pairs of system rankings based on different evaluation measures.

	2008				2009			
	P@10	P@20	R-prec.	Rel ret	P@10	P@20	R-prec.	Rel ret
MAP	0.725	0.797	0.917	0.602	0.808	0.853	0.868	0.538
P@10		0.675	0.715	0.505		0.807	0.777	0.424
P@20			0.779	0.533			0.810	0.489
R-prec.				0.589				0.466

lated, whereas their correlation with the number of relevant images retrieved is relatively low. The highest correlation is between R–precision and MAP; this has also been observed in the analysis of the TREC–7 ad hoc results (Buckley and Voorhees, 2005). Even though R–precision evaluates at exactly one point in a retrieval ranking, while MAP represents the entire area underneath the recall–precision curve, the fact that these two measures rank runs in a similar manner supports the consideration of R–precision as an overall system performance measure.

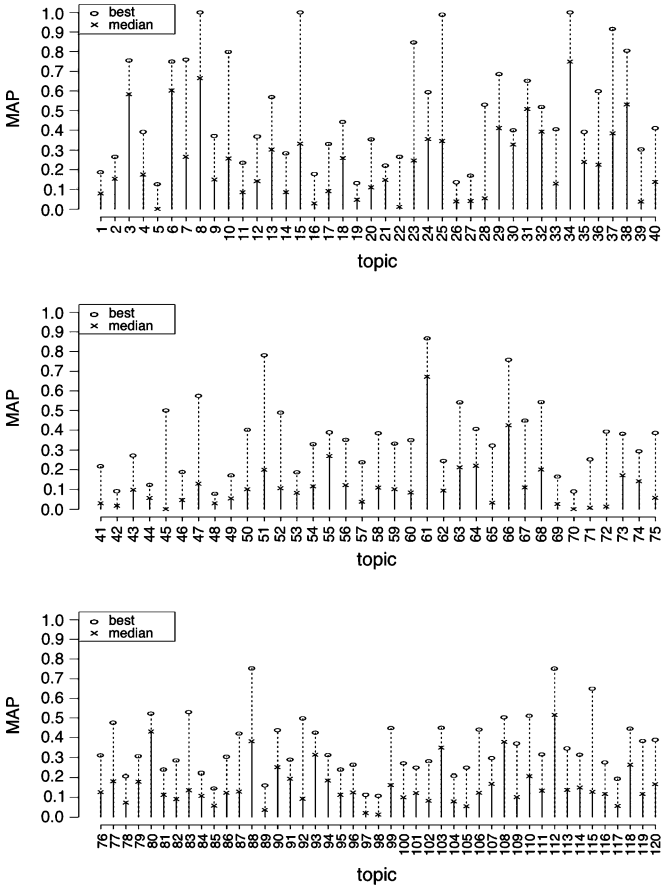


Fig. 9.5: Best and median MAP value achieved for the 2008 topics (top, middle) and for the 2009 topics (bottom).

Apart from the performance achieved over all topics in a year, it is also useful to examine the per topic performance so as to identify the problematic cases. Figure 9.5 presents for each of the topics the best MAP value achieved for that topic by a submitted run, as well as the median performance of all runs for that topic. The variability of the systems' performances over topics indicates the differences in their levels of difficulty, with some topics being very difficult for many of the submitted runs, as illustrated by the low values of the median performance. More detailed per topic analyses can be found in the overview papers of the task (Tsikrika and Kludas, 2009, 2010). Next, the results achieved by specific approaches are further examined so as to identify best practices and discuss open issues.

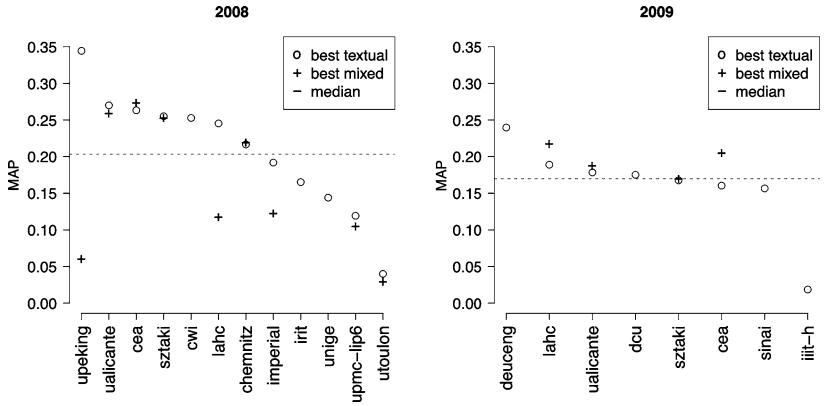


Fig. 9.6: Best textual and best mixed run (if any) for each of the participants in the Wikipedia image retrieval 2008 and 2009 tasks. The groups are ranked in decreasing order of the MAP of their best textual run.

9.4 Discussion

9.4.1 Best Practices

Over the course of these two years, a variety of different approaches have been evaluated using the test collections constructed in the context of the activities of the Wikipedia image retrieval task. To identify some of the best practices among the various techniques that have been applied, the relative performance of the submitted runs is examined.

Figure 9.6 presents for each of the groups that participated in each of the two years, the MAP achieved by its best textual and by its best mixed run (if any), together with the median MAP of all the runs submitted in that year. The group that performed best in each of the two years, *upeking* (Zhou et al, 2009) in 2008 and *deuceng* (Kilinc and Alpkocak, 2009) in 2009, applied textual query expansion using semantic knowledge bases, such a WordNet or knowledge bases extracted from Wikipedia. A similar approach was also applied by the group that achieved the third highest performance of a textual run in 2008, i.e. *cea* (Popescu et al, 2009). Furthermore, the best performing group in 2009, *deuceng* (Kilinc and Alpkocak, 2009), also applied document expansion using semantic knowledge bases. Document and query expansion using DBpedia were also applied by *dcu* (Min et al, 2010) in 2009 and achieved improvements over their textual baseline. All this constitutes strong evidence that such expansion techniques, particularly when applied judiciously so as to deal with the noise that can be potentially added, are particularly effective for

such collections of images that are accompanied by short and possibly noisy textual descriptions.

An interesting observation regarding the relative performance of textual and mixed runs is that in 2009 the groups that submitted both textual and mixed runs achieved their best results with their mixed runs. Notable cases are the *lahc* (Moulin et al, 2009, 2010) and *cea* (Popescu et al, 2009; Myoupo et al, 2010) groups that also managed to dramatically improve the performance of their mixed runs in comparison to their 2008 submissions. The improvements achieved by *lahc* were mainly due to the extraction of additional low–level visual features, including SIFT descriptors, and the combination taking place at the post–retrieval stage, as a linear combination of the text-based and visual-based rankings, rather than by considering vectors of both textual and visual terms, as they did in 2008. For *cea*, the major improvement was derived from the employment of a query model that was built using a large number of sample images automatically retrieved from the Web; in their post–submission runs, they managed to further improve the performance of their mixed runs after correcting a bug (Myoupo et al, 2010).

A final source of evidence that has also shown to be useful in this Wikipedia setting corresponds to the image names. Approaches that take them into account have shown improvements over equivalent approaches that do not in three separate cases: *ualicante* (Navarro et al, 2008) and *irit* (Torjmen et al, 2009) in 2008, and *dcu* (Min et al, 2010).

9.4.2 Open Issues

The results presented provide some clear indications on the usefulness of particular textual techniques in the context of this task but do not yet provide sufficient evidence on the best practice to follow when combining multiple modalities; further research is needed in this direction. Furthermore, apart from the encouraging results achieved in 2008 by *cea* (Popescu et al, 2009), the effectiveness of using visual concepts in an ad hoc retrieval task has not been fully explored. To this end, an effort should be made to provide classification scores for the images in the Wikipedia collection. Given the poor generalisation of concept classifiers to domains other than their training domain (Yang and Hauptmann, 2008), it would be best to build classifiers using training samples from Wikipedia. This could potentially be explored in synergy with the image annotation task (see Chapter 11 in this volume). Finally, there should be further efforts in lowering the threshold for the participation in the benchmark by providing resources to support the participants' experiments.

9.5 Conclusions and the Future of the Task

The Wikipedia image retrieval task provides test collections with the aim of supporting the reliable benchmarking of the performance of retrieval approaches that exploit textual and visual evidence for ad hoc image retrieval in the context of a large and heterogeneous collection of freely distributable Wikipedia images that are searched for by users with diverse information needs. Over the course of these two years at ImageCLEF, a variety of retrieval approaches have been been investigated and interesting conclusions have been reached regarding best practices in the field. Nonetheless, much work remains to be done. Future runs of the task will continue to examine the same evaluation objectives using even larger image collections (already the collection provided in 2010 consists of approximately 250,000 Wikipedia images) and exploring their multi–lingual aspects. Further experimentation with the test collections constructed thus far is possible by downloading them from ImageCLEF's resources page[10].

Acknowledgements Theodora Tsikrika was supported by the European Union via the European Commission project VITALAS (EU–IST–IP 045389). Jana Kludas was funded by the European Commission project MultiMATCH (EU–IST–STREP 033104) and the Swiss National Fund (SNF). The authors would also like to thank Thijs Westerveld and Roelof van Zwol for creating the (INEX MM) Wikipedia image collection and setting up the task in the context of the activities of the INEX Multimedia track, Thomas Deselaers for invaluable technical support during ImageCLEF 2008, and all the groups that participated in the task and in the relevance assessment process.

References

Arni T, Clough PD, Sanderson M, Grubinger M (2009) Overview of the ImageCLEFphoto 2008 photographic retrieval task. In: Peters et al (2009), pp 500–511

Buckley C, Voorhees EM (2005) Retrieval system evaluation. In: Voorhees EM, Harman DK (eds) TREC: Experiment and Evaluation in Information Retrieval, Digital Libraries and Electronic Publishing, MIT Press, chap 3, pp 53–75

Daróczy B, Petrás I, Benczúr AA, Fekete Z, Nemeskey D, Siklósi D, Weiner Z (2009) SZTAKI @ ImageCLEF 2009. In: Working Notes of CLEF 2009, Corfu, Greece

Denoyer L, Gallinari P (2007) The Wikipedia XML corpus. In: Fuhr et al (2007), pp 12–19

Díaz-Galiano M, Martín-Valdivia M, Urena-López L, Perea-Ortega J (2010) Using WordNet in multimedia information retrieval. In: Peters et al (2010)

Fakeri-Tabrizi A, Amini MR, Tollari S, Gallinari P (2008) UPMC/LIP6 at ImageCLEF wikipediaMM: an image–annotation model for an image search–engine. In: Working Notes of CLEF 2008, Aarhus, Denmark

Ferecatu M (2005) Image retrieval with active relevance feedback using both visual and keyword-based descriptors. In: Ph.D. Thesis, Université de Versailles, France

Fuhr N, Lalmas M, Trotman A (eds) (2007) Comparative Evaluation of XML Information Retrieval Systems, Proceedings of the 5th International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2006), Revised Selected Papers, Lecture Notes in Computer Science (LNCS), vol 4518, Springer

[10] http://www.imageclef.org/datasets/

van Gemert JC, Geusebroek JM, Veenman CJ, Snoek CGM, Smeulders AWM (2006) Robust scene categorization by learning image statistics in context. In: Proceedings of the 2006 Conference on Computer Vision and Pattern Recognition Workshop. IEEE Computer Society, Washington, DC, USA, p 105

Grubinger M, Clough PD, Leung C (2006) The IAPR TC–12 benchmark for visual information search. IAPR Newsletter 28(2):10–12

Kilinc D, Alpkocak A (2009) DEU at ImageCLEF 2009 wikipediaMM task: Experiments with expansion and reranking approaches. In: Working Notes of CLEF 2009, Corfu, Greece

Kludas J, Tsikrika T (2008) ImageCLEF 2008 wikipediaMM task guidelines. Unpublished document distributed to ImageCLEF 2008 wikipediaMM participants

Larsen B, Trotman A (2006) INEX 2006 guidelines for topic development. In: Fuhr N, Lalmas M, Trotman A (eds) Preproceedings of the 5th International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2006), pp 373–380

Malik S, Trotman A, Lalmas M, Fuhr N (2007) Overview of INEX 2006. In: Fuhr et al (2007), pp 1–11

Min J, Wilkins P, Leveling J, Jones GJF (2010) Document expansion for text-based image retrieval at CLEF 2009. In: Peters et al (2010)

Moulin C, Barat C, Géry M, Ducottet C, Largeron C (2009) UJM at ImageCLEFwiki 2008. In: Peters et al (2009), pp 779–786

Moulin C, Barat C, Lemaître C, Géry M, Ducottet C, Largeron C (2010) Combining text/image in wikipediaMM task 2009. In: Peters et al (2010)

Myoupo D, Popescu A, Borgne HL, Moëllic PA (2010) Multimodal image retrieval over a large database. In: Peters et al (2010)

Navarro S, Muñoz R, Llopis F (2008) A textual approach based on passages using IR–n in wikipediaMM task 2008. In: Working Notes for the CLEF 2008 Workshop, Aarhus, Denmark

Navarro S, Muñoz R, Llopis F (2009) Evaluating fusion techniques at different domains at ImageCLEF subtasks. In: Working Notes of CLEF 2009, Corfu, Greece

Overell S, Llorente A, Liu H, Hu R, Rae A, Zhu J, Song D, Rüger S (2008) MMIS at ImageCLEF 2008: Experiments combining different evidence sources. In: Working Notes for the CLEF 2008 Workshop, Aarhus, Denmark

Peters C, Deselaers T, Ferro N, Gonzalo J, Jones GJF, Kurimo M, Mandl T, Peñas A, Petras V (eds) (2009) Evaluating Systems for Multilingual and Multimodal Information Access: Proceedings of the 9th Workshop of the Cross–Language Evaluation Forum (CLEF 2008), Revised Selected Papers, Lecture Notes in Computer Science (LNCS), vol 5706, Springer

Peters C, Tsikrika T, Müller H, Kalpathy-Cramer J, Jones GJF, Gonzalo J, Caputo B (eds) (2010) Multilingual Information Access Evaluation II, Multimedia Experiments: Proceedings of the 10th Workshop of the Cross–Language Evaluation Forum (CLEF 2009), Revised Selected Papers, Lecture Notes in Computer Science (LNCS), Springer

Popescu A, Borgne HL, Moëllic PA (2009) Conceptual image retrieval over a large scale database. In: Peters et al (2009), pp 771–778

Racz S, Daróczy B, Siklósi D, Pereszlényi A, Brendel M, Benczúr AA (2008) Increasing cluster recall of cross-modal image retrieval. In: Working Notes for the CLEF 2008 Workshop, Aarhus, Denmark

Snoek CGM, Worring M, van Gemert JC, Geusebroek JM, Smeulders AWM (2006) The challenge problem for automated detection of 101 semantic concepts in multimedia. In: Proceedings of the 14th Annual ACM International Conference on Multimedia. ACM press, New York, NY, USA, pp 421–430

Torjmen M, Pinel-Sauvagnat K, Boughanem M (2009) Evaluating the impact of image names in context-based image retrieval. In: Peters et al (2009), pp 756–762

Tsikrika T, Kludas J (2009) Overview of the wikipediaMM task at ImageCLEF 2008. In: Peters et al (2009), pp 539–550

Tsikrika T, Kludas J (2010) Overview of the wikipediaMM task at ImageCLEF 2009. In: Peters et al (2010)

Tsikrika T, Westerveld T (2008) The INEX 2007 Multimedia track. In: Fuhr N, Lalmas M, Trotman A, Kamps J (eds) Focused access to XML documents, Proceedings of the 6th International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2007), Springer, Lecture Notes in Computer Science (LNCS), vol 4862, pp 440–453

Tsikrika T, Rode H, de Vries AP (2008) CWI at ImageCLEF 2008. In: Working Notes for the CLEF 2008 Workshop, Aarhus, Denmark

Vundavalli S (2009) IIIT-H at ImageCLEF Wikipedia MM 2009. In: Working Notes of CLEF 2009, Corfu, Greece

Westerveld T, van Zwol R (2007) The INEX 2006 Multimedia track. In: Fuhr et al (2007), pp 331–344

Wilhelm T, Kürsten J, Eibl M (2008) The Xtrieval framework at CLEF 2008: ImageCLEF wikipediaMM task. In: Working Notes for the CLEF 2008 Workshop, Aarhus, Denmark

Yang J, Hauptmann AG (2008) (Un)Reliability of video concept detection. In: Luo J, Guan L, Hanjalic A, Kankanhalli MS, Lee I (eds) Proceedings of the 7th International Conference on Content–based Image and Video Retrieval (CIVR 2008), ACM press, pp 85–94

Zhao ZQ, Glotin H (2008) Concept content based Wikipedia web image retrieval using CLEF VCDT 2008. In: Working Notes of CLEF 2008, Aarhus, Denmark

Zhou Z, Tian Y, Li Y, Huang T, Gao W (2009) Large–scale cross-media retrieval of wikipediaMM images with textual and visual query expansion. In: Peters et al (2009), pp 763–770