

Chapter 5

Task Creation and Visual Information Complexity

The creation of the *IAPR TC-12 image collection* (see previous chapter) has certainly provided one of the key foundations for the successful organisation of an evaluation event for VIR from generic photographic collections (*i.e.* containing everyday real-world photographs akin to those that can frequently be found in private photographic collections as well, *e.g.* holiday pictures or photos of sporting events). Based on this novel resource, the design of the next, no less crucial, benchmark component can be commenced: the creation of retrieval tasks and their corresponding query topics, which represents another key aspect of a test collection as one must aim to create a balanced and representative set of information needs.

In practice, this is achieved by generating topics against certain dimensions, including the opinion of domain experts, the estimated number of relevant documents (images) for each topic, the topic scope (*e.g.* broad or narrow, general or specific) and a variation of additional task parameters such as geographic constraints (see Section 3.3.2). Still, these search topics were often not considered as representative for real-world information needs [120] and sometimes as either too difficult (and occasionally even too easy) for the state-of-the-art image retrieval methods.

In this chapter, we take on these two issues in order to steal the critics' thunder and to mitigate their arguments. First, we introduce a novel measure to quantify and control the retrieval difficulty of concept-based image queries: Section 5.1 moti-

vates and summarises related work in quantifying task and topic difficulty; Section 5.2 presents its definition and examples; and Section 5.3 investigates the accuracy of our novel measure based on its correlation with system effectiveness.

Then, Section 5.4 describes the model that we established in order to facilitate the *topic creation process* for image retrieval evaluation events; it reports on the results of a log file analysis that we carried out to form a pool of realistic and representative candidate topics for our specific collection. Based on these topic candidates, we created a set of representative search queries against a number of dimensions which also included the new difficulty measure.

We finally conclude by listing the benefits and contributions achieved in this chapter.

5.1 Introduction to Retrieval Difficulty

Research in information retrieval (IR) has recently focused on estimating the difficulty of a query (*i.e.* the difficulty for retrieval systems to return relevant images): an attempt to quantify the quality of search results [38]. Most of this section is taken from [149].

5.1.1 Motivation

Being able to estimate the difficulty of a query has appealing benefits for both the individual and organisations. For example, users of an IR system could be shown how well their query is likely to perform; or search engine companies could identify topics of interest to their users, which are not being answered well by the system [37].

A further use of estimating the difficulty of a search query is to help select suitable search topics for evaluation events. It is crucial to balance such topics for difficulty: they should be neither too easy nor too hard. As image retrieval algorithms improve, it is necessary to increase the average difficulty level of topics each year in order to maintain the challenge for returning participants. However, if

topics are too difficult for the existing techniques, the results are not particularly meaningful. Furthermore, it may prove difficult for new participants to obtain decent results and prevent them from presenting their findings and taking part in comparative evaluations. On the other hand, if topics are too easy, participants can achieve good results without using sophisticated approaches, which might slow down the research progress and make the ranking of systems hard.

A benchmark should therefore ideally exhibit topics with a range of difficulty levels; and being able to quantify topic difficulty has obvious benefits for both the organisers of the evaluation campaign and participants in allowing them to observe (and analyse) retrieval effectiveness with respect to topic difficulty levels.

Hence, in this chapter, we consider the problem of estimating topic difficulty for TBIR to assist with the topic selection process during benchmark creation. The notion of topic difficulty investigated here is based on linguistic (and statistical) features which might affect the successful retrieval of images from the retrieval system's (and not the user's) point of view. That is, we assume that effective (or good) retrieval by an IR system is reflected by the value of measures such as *MAP* and *P(10)*.

The level of correlation with system effectiveness can then be used to indicate how effective the measure is at estimating topic difficulty: we assume that “difficult” topics will result in a lower *MAP* or *P(10)*, and this is akin to the view held by most previous research on topic difficulty [37].

5.1.2 Query Difficulty

Quantifying task difficulty is not a new concept; on the contrary, it has been applied to many areas such as machine learning [321], parsing and grammatical formalisms [22], and language learning in general [356], while early papers in the field of IR include domain complexity with respect to information extraction tasks [15], the discussion of syntactic complexity in multimedia information retrieval [116], and a measure of semantic complexity for natural language systems [345]. More recent

papers include measures of topic difficulty for information retrieval [74, 500]. In the context of the research presented in this chapter, the most relevant of these are: syntactic complexity in multimedia IR and topic difficulty in IR.

Related Work

Flank [116] showed that TBIR based on full-sentence captions performs better than retrieval using captions composed of word lists. The use of natural language processing (NLP) techniques with IR on images annotated with grammatical units (a system called *PictureQuest*) produces higher retrieval accuracy than standard IR on word lists alone.

In information retrieval, Cronen-Townsend et al. [74] introduced the *clarity score* as an attempt to estimate whether a query was going to be difficult. The score measures the difference between the query language model and the corresponding document language model and shows positive rank correlations (Spearman) between 0.39 and 0.58 with the average precision of the topics in several TREC collections. The *divergence from randomness (DFR)* scoring model [7] also claims to show a positive correlation of 0.52 with query precision.

Recent Development

Recent events such as the *Reliable Information Access (RIA)* workshop [168], the *Robust Retrieval Track* of TREC [470] and the *SIGIR 2005 workshop on predicting query difficulty* [38] have generated interest and discussion on topic difficulty. It has been shown that IR systems perform worse on more difficult topics and that being able to recognise factors contributing to topic difficulty could help improve IR retrieval accuracy.

Approaches for estimating query difficulty resulting from these events have shown that features correlated with difficulty include the frequency of document terms in the collection [221, 102], the linguistic composition of the query [286], the coherence of relevant documents [108], and the agreement between the top results of the full query and the top results of its sub-queries [500]. The latter approach

reports rank correlation scores (Kendall) of up to 0.57 with the MAP of topics from TREC-8.

Mothe and Tanguy [286] consider sixteen linguistic features and their correlation with TREC average precision scores. They find that horizontal syntactic complexity (syntactic links span) and semantic ambiguity (a polysemy value) correlate most strongly with system effectiveness: the highest product-moment correlation (Pearson) achieved was -0.40.

Carmel *et al.*[37] provide an approach which models the relationship between topic texts, the set of relevant documents and the collection, and they show empirically that topic difficulty correlates strongly with the distances between these components, with correlation values (Pearson) between 0.45 and 0.48.

5.1.3 Topic Difficulty in Benchmarks

While quantifying task difficulty is not a totally new concept in the field of VIR, little work has considered topic difficulty as a dimension for the topic development process (Eguchi *et al.*, for example, investigated the topic difficulty for *NTCIR* [102]).

It is hereby important to note that the prediction of query difficulty (*e.g.* clarity scores, convergence from randomness, *etc.*) and the estimation of topic difficulty are not looking at the same problem, because topics are not easy or difficult in isolation, but depend on the document collection to be searched. A comparison of measures for query and topic difficulty is therefore not meaningful, as there can be bad queries for easy topics, for example, and vice versa.

No work has considered topic difficulty for concept-based image retrieval benchmarks, which is one of the major contributions described in this chapter. We therefore designed a novel measure for topic difficulty in concept-based image retrieval, which is defined and validated in Sections 5.2 and 5.3 hereinafter.

5.2 Topic Difficulty Model

This section explains the model for our topic difficulty measure for concept-based image retrieval benchmarks. Similar to previous work, we consider a topic’s difficulty to be influenced by its linguistic composition and the statistical relationship of query terms with the document collection and set of relevant documents. However, since the focus of this work is to assist with topic selection, the approach described here involves more manual effort than those previously reported (but given that the process of topic generation is manual, this is not an unrealistic assumption).

Our measure uses syntactic complexity [116] and grammatical sentence elements [149] as the basis for linguistic analysis, rather than individual query terms (Section 5.2.1). In addition, a factor is included which allows for the differences between the visual contents of an image and the corresponding semantic description (the *annotation gap* - Section 5.2.2). Finally, the difficulty score is computed using an iterative calculation based on the most significant topic element at each step (Section 5.2.3).

To arrive at this algorithm, we tested 20 different approaches (see Section 5.3.2), with the one described hereinafter showing the most promising results.

5.2.1 Sentence Elements

This approach is based on analysing the grammatical sentence elements of a topic, not individual query terms. We tested both approaches (see Section 5.3.2), but achieved a higher correlation with grammatical analysis based on sentence elements, which also corresponds with the findings of previous studies such as [116, 149].

Definition 5.2.1 *Let t be a grammatical sentence element. We say that t is a topic sentence element if*

$$t \in \{\textit{numerals}, \textit{nouns}, \textit{verbs}, \textit{adjectives}, \textit{adverbs}, \textit{adjuncts}\} \quad (5.1)$$

where adjuncts are either locative, temporal, causal or modicative adjuncts.

Definition 5.2.2 Given a sentence T , we define the topic sentence \mathcal{T} as

$$\mathcal{T} := \{t_1, t_2, \dots, t_K\} \quad (5.2)$$

where t_k denotes the k^{th} topic sentence element¹ of T , K is the number of topic sentence elements in T , and $1 \leq k \leq K$.

Definition 5.2.3 Let t_k be the k^{th} topic sentence element of a topic sentence \mathcal{T} as defined in Definition 5.2.2. We denote \mathcal{R}_k as the set of all relevant images for topic sentence element t_k , and define \mathcal{R} as

$$\mathcal{R} := \{\mathcal{R}_1, \mathcal{R}_2, \dots, \mathcal{R}_K\}, \quad (5.3)$$

where K is the number of topic sentence elements in \mathcal{T} , and $1 \leq k \leq K$.

5.2.2 Quantifying the Annotation Gap

The measure being defined is for topic difficulty in concept-based *image retrieval*. This is different from standard document retrieval as one must also consider the distance between the alphanumeric image representations (which are considered as relevant) and the set of relevant images themselves.

This distance (which we refer to as *annotation gap*) can be due to at least two different reasons. Firstly, an image retrieval algorithm based on text only may not be able to return all relevant images because of vocabulary mismatches and incomplete, wrong or missing captions (annotations). Consider the query term “people”: due to vocabulary mismatch, not all relevant images may be returned because some may exhibit the use of synonyms or hypernyms (*e.g.* men, women, children, spectators); some images may be annotated incorrectly (*e.g.* with typographical errors like “peolpe”); others may have incomplete (or sparse) semantic representations (*e.g.* not containing the term or its variations) due to the lack of associated text.

¹We only consider the stems (or roots) of the topic elements rather than full words hereinafter, so that matches are not missed through trivial word variations later on (*e.g.* differentiation between singular and plural forms, *etc.*).

Definition 5.2.4 Let \mathcal{R} denote the set of relevant images for a query term, and \mathcal{N}_D the set of images retrieved as direct hits for that query term (i.e. images that are found directly through their semantic descriptions). We define the factor for vocabulary mismatch and incomplete or incorrect annotation α as

$$\alpha := 1 - \frac{|\mathcal{N}_D \cap \mathcal{R}|}{|\mathcal{R}|} \quad (5.4)$$

with $|\cdot|$ denoting the cardinality of a set.

The definition of α in (5.4) implies that $0 \leq \alpha \leq 1$. If $\alpha > 0$, then only some of the relevant images are automatically retrieved.

Example 5.2.1 We assume a data collection that contains 5000 images that show the topic term *people* ($|\mathcal{R}| = 5000$). We further assume that 3000 of these images are directly retrieved because they are annotated with the term “*people*” ($|\mathcal{N}_D| = 3000$), while a further 1000 images that show *people* are indirectly annotated with synonyms or hypernyms thereof (e.g. *men*, *women*, *children*, *crowd*, *spectators*, etc.), and that the remaining 1000 images that show *people* are not annotated as such. Then

$$\alpha = 1 - \frac{|\mathcal{N}_D \cap \mathcal{R}|}{|\mathcal{R}|} = 1 - \frac{3000}{5000} = 0.4.$$

The second reason for the annotation gap is as follows: a concept-based retrieval algorithm may return *more* images than are relevant due to incorrect annotation and word (or sentence element) ambiguity. Consider a query for the Californian city of “San Francisco”: not all the images returned that contain the term “San Francisco” are relevant due to incorrect senses (e.g. South American cathedrals called “San Francisco”), incorrect annotation (e.g. an image of Los Angeles incorrectly annotated as “San Francisco”), and language-specific translations (e.g. the Spanish “San Francisco” could retrieve images of the Catholic saint “Francis of Assisi”).

Definition 5.2.5 Given \mathcal{R} the set of relevant images for a topic term, and \mathcal{N} the set of images retrieved, we define the factor for element ambiguity β as

$$\beta := 1 - \frac{|\mathcal{R}|}{|\mathcal{N} \cup \mathcal{R}|} \quad (5.5)$$

The definition of β in (5.5) implies that $0 \leq \beta \leq 1$. If $\beta > 0$, then only some of the images that are retrieved as relevant are, in fact, relevant.

Example 5.2.2 We assume that 1000 images of a collection are annotated with “San Francisco”: 400 of them show the Californian city, and 600 show churches from South America called “San Francisco”. In a query topic for the Californian city, 1000 images would be returned ($|\mathcal{N}| = 1000$), although only 400 are, in fact, relevant ($|\mathcal{R}| = 400$). Thus,

$$\beta = 1 - \frac{|\mathcal{R}|}{|\mathcal{N} \cup \mathcal{R}|} = 1 - \frac{400}{1000} = 0.6.$$

Definition 5.2.6 Let α and β be as defined in Definition 5.2.4 and Definition 5.2.5 respectively. We define the annotation gap factor γ as

$$\gamma := \eta + [\theta\alpha + (1 - \theta)\beta] \quad (5.6)$$

where $\eta \in \mathbb{R}$ and $0 \leq \theta \leq 1$.

Empirical investigation using two image collections (*SAC* and the *IAPR TC-12* image collection), 113 topics and the results from three *ImageCLEF* campaigns (2004 - 2006) has shown that the highest correlation with the average MAP values is obtained using the parameter values $\eta = 1.2$ and $\theta = 0.6$ (see Figure 5.1).

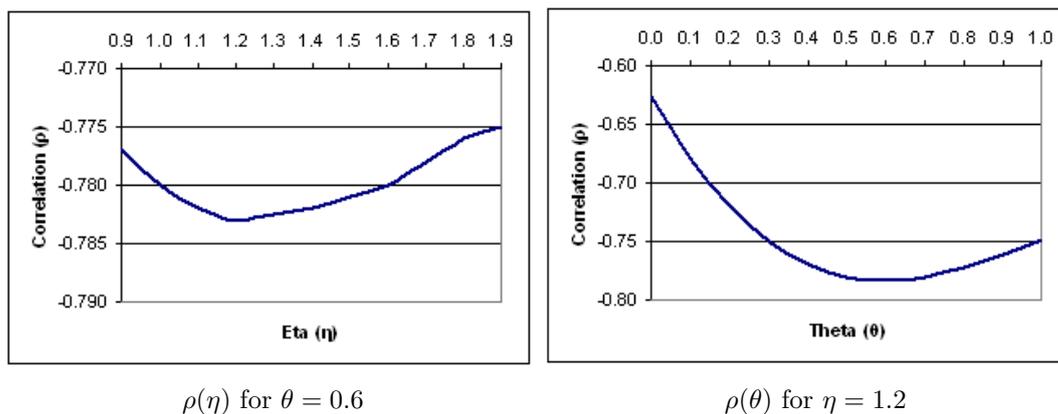


Figure 5.1: Empirical investigations for η and θ .

5.2.3 Topic Difficulty

Based on the definitions in Section 5.2.2, we compute the difficulty $d(\mathcal{T}, \mathcal{I})$ for a topic sentence \mathcal{T} and an image collection \mathcal{I} by calculating the sum of the topic difficulty for each of the iterations as follows.

Initial condition. We denote $\mathcal{R}_{0,k}$ ($k \in \{1, \dots, K\}$) as the set of relevant images for topic sentence element $t_k \in \mathcal{T}$, and define

$$\mathcal{R}_0 := \{\mathcal{R}_{0,1}, \mathcal{R}_{0,2}, \dots, \mathcal{R}_{0,K}\}. \quad (5.7)$$

We also denote $\mathcal{R}_0^* = \mathcal{I}$ and $\mathcal{R}_0 = \mathcal{R}$ (see Definition 5.2.3).

Iteration 1. We denote \mathcal{R}_1 as the set that contains all the sets of relevant images for the first iteration, that is,

$$\mathcal{R}_1 := \{\mathcal{R}_{1,1}, \mathcal{R}_{1,2}, \dots, \mathcal{R}_{1,K}\}, \quad (5.8)$$

where $\mathcal{R}_{1,k} = \mathcal{R}_{0,k}$. Now, we define

$$\mathcal{R}_1^* := \mathcal{R}_{1,l},$$

where $l = l(1)$ is the index where the cardinality of the sets $\mathcal{R}_{1,k}$ ($k \in \{1, \dots, K\}$) attains its minimum, that is,

$$|\mathcal{R}_{1,l}| = \min_k (|\mathcal{R}_{1,k}|).$$

Next, we calculate the linear, conditional document frequency of the set of relevant images $\mathcal{R}_{1,k}$ for the k^{th} topic element,

$$df(\mathcal{R}_{1,k}) = P(\mathcal{R}_{1,k} | \mathcal{R}_0^*) = \frac{P(\mathcal{R}_{1,k} \cap \mathcal{R}_0^*)}{P(\mathcal{R}_0^*)} = \frac{|\mathcal{R}_{1,k}|}{|\mathcal{R}_0^*|}, \quad (5.9)$$

(here, we interpret the frequency as probability and denote it by P). Then, we calculate d_1 as the difficulty for the most significant element of the first iteration

$$d_1 := [1 - df(\mathcal{R}_1^*)] \gamma_1^*, \quad (5.10)$$

where γ_1^* is the annotation gap factor for the most significant element set of the first iteration.

Iteration j. To build \mathcal{R}_j , *i.e.* the set that contains all the sets of relevant images for the j^{th} iteration, we only consider $(K - j + 1)$ topic elements by deleting the index $l = l(j - 1)$ (*i.e.* the index where the cardinality of the sets $\mathcal{R}_{j-1,k}$ attains its minimum) and left-shifting the remaining indices $k \geq l + 1$ (*i.e.* shifting k to $k - 1$):

$$\mathcal{R}_j := \{\mathcal{R}_{j,1}, \mathcal{R}_{j,2}, \dots, \mathcal{R}_{j,K-j+1}\}, \quad (5.11)$$

where

$$\mathcal{R}_{j,k} := \mathcal{R}_{j-1}^* \cap \mathcal{R}_{j-1,k}. \quad (5.12)$$

Now, we define

$$\mathcal{R}_j^* := \mathcal{R}_{j,l}, \quad (5.13)$$

where $l = l(j)$ is the index where the cardinality of the sets $\mathcal{R}_{j,k}$ ($k \in \{1, \dots, K - j + 1\}$) attains its minimum, that is,

$$|\mathcal{R}_{j,l}| = \min_k (|\mathcal{R}_{j,k}|).$$

Next, we calculate the linear, conditional document frequency of the set of relevant images $\mathcal{R}_{j,k}$ for the k^{th} topic element,

$$df(\mathcal{R}_{j,k}) = P(\mathcal{R}_{j,k} | \mathcal{R}_{j-1}^*) = \frac{P(\mathcal{R}_{j,k} \cap \mathcal{R}_{j-1}^*)}{P(\mathcal{R}_{j-1}^*)} = \frac{|\mathcal{R}_{j,k}|}{|\mathcal{R}_{j-1}^*|}. \quad (5.14)$$

Then, we calculate d_j as the difficulty for the most significant element of the j^{th} iteration

$$d_j := [1 - df(\mathcal{R}_j^*)] \gamma_j^*, \quad (5.15)$$

where γ_j^* is the annotation gap factor for the most significant element set of the j^{th} iteration.

Final result. Finally, we calculate $d = d(\cdot, \cdot)$, *i.e.* the difficulty for a topic sentence \mathcal{T} and an image collection \mathcal{I} , as

$$d := \sum_{j=1}^K d_j. \quad (5.16)$$

5.2.4 Examples

This section provides a detailed sample elaboration on the calculation of the difficulty for the topics “people in San Francisco” and “photos of female guides” in order to clearly depict the algorithm introduced above.

Example 1: People in San Francisco

Given are the topic sentence “people in San Francisco”, an image collection \mathcal{I} that contains a total of 20,000 images ($N = |\mathcal{I}| = 20,000$), and the cardinality values already used in the examples in Section 5.2.2.

Initial condition. We first define the topic sentence \mathcal{T} which contains two topic sentence elements: the noun *people* (t_1) and the locative adjunct *San Francisco* (t_2), whereby we remove the preposition (*in*) because it is found in most stop-word lists:

$$\mathcal{T} = \{people, SanFrancisco\}.$$

In addition, we also define the set of relevant images (\mathcal{R}) for each of the topic sentence elements, whereby \mathcal{R}_1 is the set of all the images containing *people* (t_1) and \mathcal{R}_2 the set of all the images that were taken in *San Francisco* (t_2). Being the starting point of the algorithm, we denote \mathcal{R} as the zeroth iteration (\mathcal{R}_0):

$$\mathcal{R}_0 = \{\mathcal{R}_{0,1}, \mathcal{R}_{0,2}\}.$$

Since there are two sentence elements in the topic sentence ($K = 2$), we have to go through two iterations in order to arrive at the final topic difficulty value.

Iteration 1. In the first iteration ($j = 1$), we have $\mathcal{T}_1 = \mathcal{T}$ and $\mathcal{R}_{1,k} = \mathcal{R}_{0,k}$ by definition, and therefore

$$\mathcal{T}_1 = \{people, SanFrancisco\}.$$

$$\mathcal{R}_1 = \{\mathcal{R}_{1,1}, \mathcal{R}_{1,2}\}.$$

We first compare the cardinalities of both sets $\mathcal{R}_{1,1}$ and $\mathcal{R}_{1,2}$: there are 5000 images in the database that contain people ($|\mathcal{R}_{1,1}| = 5000$) and 400 images that were taken in San Francisco ($|\mathcal{R}_{1,2}| = 400$). The most significant topic element is therefore *San Francisco*, because it contains the minimum number of relevant images in the first iteration:

$$\mathcal{R}_1^* = \mathcal{R}_{1,2}.$$

Next, we calculate the linear topic frequency of this element, bearing in mind that in the first iteration, the most significant element of the “previous” iteration is the entire image database by default, $\mathcal{R}_0^* = \mathcal{I}$, and therefore

$$df(\mathcal{R}_1^*) = \frac{|\mathcal{R}_{1,2}|}{|\mathcal{R}_0^*|} = \frac{|\mathcal{R}_{1,2}|}{|\mathcal{I}|} = \frac{|\{\text{images of San Francisco}\}|}{|\{\text{all images in the collection}\}|} = \frac{400}{20000} = 0.02.$$

Then, we compute the annotation gap factor for the most significant element of the first iteration (γ_1^*) using the values that we have already determined in Section 5.2.2 ($\alpha = 0$, $\beta = 0.6$, $\eta = 1.2$ and $\theta = 0.6$):

$$\gamma_1^* = \eta + [\theta\alpha_1^* + (1 - \theta)\beta_1^*] = 1.2 + [0.6 * 0.4 + 0.4 * 0.6] = 1.68.$$

Finally, we can calculate the topic difficulty d_1 for the first iteration:

$$d_1 = [1 - df(\mathcal{R}_1^*)]\gamma_1^* = [1 - 0.02]1.68 = 1.65.$$

Iteration 2. In the second iteration ($j = 2$), we first have to build \mathcal{T}_2 and \mathcal{R}_2 by intersecting the set of all relevant images of the most significant topic element of the previous iteration \mathcal{R}_1^* with all the sets of relevant images of the previous iteration $\mathcal{R}_{1,k}$, except $\mathcal{R}_1^* = \mathcal{R}_{1,2}$ itself:

$$\mathcal{T}_2 = \{(people, SanFrancisco)\},$$

$$\mathcal{R}_2 = \{\mathcal{R}_{2,1}\} = \{\mathcal{R}_1^* \cap \mathcal{R}_{1,1}\}.$$

There is only one sentence element left, which is automatically the most significant element for the second iteration:

$$\mathcal{R}_2^* = \mathcal{R}_{2,1}.$$

We now assume that 200 images show *people* in the Californian city of *San Francisco* ($|\mathcal{R}_2^*| = 200$), that 150 of them are found by direct hits ($|\mathcal{N}_{D_2^*} \cap \mathcal{R}_2^*| = 150$), which means they are directly annotated with people in San Francisco (and 50 indirectly with men, women, children in San Francisco), and that a total of 300 images show people in the Californian city of San Francisco as well as in South American churches called San Francisco ($|\mathcal{N}_2^* \cup \mathcal{R}_2^*| = 300$). The linear document frequency df therefore is

$$df(\mathcal{R}_2^*) = \frac{|\mathcal{R}_{2,1}|}{|\mathcal{R}_1^*|} = \frac{|\{\text{images of people in San Francisco}\}|}{|\{\text{images of San Francisco}\}|} = \frac{200}{400} = 0.5,$$

the factor for vocabulary mismatch and incomplete and incorrect annotation

$$\alpha_2^* = 1 - \frac{|\mathcal{N}_{D_2^*} \cap \mathcal{R}_2^*|}{|\mathcal{R}_2^*|} = 1 - \frac{150}{200} = 1 - 0.75 = 0.25,$$

the factor for word ambiguity

$$\beta_2^* = 1 - \frac{|\mathcal{R}_2^*|}{|\mathcal{N}_2^* \cup \mathcal{R}_2^*|} = 1 - \frac{200}{300} = 1 - 0.66 = 0.33,$$

the factor for the annotation gap

$$\gamma_2^* = \eta + [\theta\alpha_2^* + (1 - \theta)\beta_2^*] = 1.2 + [0.6 * 0.25 + 0.4 * 0.33] = 1.38,$$

and the difficulty for the second iteration

$$d_2 = [1 - df(\mathcal{R}_2^*)] * \gamma_2^* = [1 - 0.5] * 1.38 = 0.69.$$

Final result. The total topic difficulty of topic \mathcal{T} (“people in San Francisco”) for an image collection \mathcal{I} after two iterations amounts to

$$d(\mathcal{T}, \mathcal{I}) = \sum_{j=1}^2 d_j = 1.65 + 0.69 = 2.34.$$

Example 2: Photos of Female Guides

While the first example was a rather contrived one in order to depict the topic difficulty algorithm as clearly as possible, this example will now elaborate on one (realistic) sample topic taken from the *ImageCLEFphoto 2006* event: given are the topic sentence “photos of female guides” and the *IAPR TC-12 image collection* \mathcal{I} that contains a total of 20,000 images ($N = |\mathcal{I}| = 20,000$).

Initial condition. We first define the topic sentence \mathcal{T} which contains three topic sentence elements: the noun *photo* (t_1), the adjective *female* (t_2) and the noun *guide* (t_3). Similar to the first example, we remove the adposition (*of*) because it is found in most stop-word lists, and we also consider the stemmed versions of the original sentence elements used in the topic sentence:

$$\mathcal{T} = \{photo, female, guide\}.$$

Again, we also define the set of relevant images (\mathcal{R}) for each of the topic sentence elements, whereby \mathcal{R}_1 is the set of all the images that are *photos* (t_1), \mathcal{R}_2 the set of all the images containing anything *female* (t_2), and \mathcal{R}_3 the set of all the images that show *guides* (t_3). Again, being the starting point of the algorithm, we denote \mathcal{R} as the zeroth iteration (\mathcal{R}_0):

$$\mathcal{R}_0 = \{\mathcal{R}_{0,1}, \mathcal{R}_{0,2}, \mathcal{R}_{0,3}\}.$$

Since there are three sentence elements in the topic sentence ($K = 3$), we have to go through three iterations in order to arrive at the final topic difficulty value.

Iteration 1. In the first iteration ($j = 1$), we have $\mathcal{T}_1 = \mathcal{T}$ and $\mathcal{R}_{1,k} = \mathcal{R}_{0,k}$ by definition, and therefore

$$\mathcal{T}_1 = \{photo, female, guide\}.$$

$$\mathcal{R}_1 = \{\mathcal{R}_{1,1}, \mathcal{R}_{1,2}, \mathcal{R}_{1,3}\}.$$

We first compare the cardinalities of the three sets $\mathcal{R}_{1,1}$, $\mathcal{R}_{1,2}$ and $\mathcal{R}_{1,3}$: all images in the *IAPR TC-12 image collection* are *photos* ($|\mathcal{R}_{1,1}| = 20000$), there are 3045 images showing something *female* ($|\mathcal{R}_{1,2}| = 3045$) and 90 images of *guides* ($|\mathcal{R}_{1,3}| = 90$). The most significant topic element is therefore *guide*, because it contains the minimum number of relevant images in the first iteration:

$$\mathcal{R}_1^* = \mathcal{R}_{1,3}.$$

Next, we calculate the linear document frequency of this element, again bearing in mind that in the first iteration, the most significant element of the “previous” iteration is the entire image database by default, $\mathcal{R}_0^* = \mathcal{I}$, and therefore

$$df(\mathcal{R}_1^*) = \frac{|\mathcal{R}_{1,3}|}{|\mathcal{R}_0^*|} = \frac{|\{\text{images of guides}\}|}{|\{\text{all images in the collection}\}|} = \frac{90}{20000} = 0.0045.$$

Then, we compute the annotation gap factor for the most significant element of the first iteration (γ_1^*). In the *IAPR TC-12 image collection*, all photos of guides are also annotated as such ($|\mathcal{N}_{D_1}^* \cap \mathcal{R}_1^*| = 90$) and there are no other photos of guides in the database that are not relevant ($|\mathcal{N}_1^* \cup \mathcal{R}_1^*| = 90$), therefore $\alpha_1^* = 0$ and $\beta_1^* = 0$, and subsequently

$$\gamma_1^* = \eta + [\theta\alpha_1^* + (1 - \theta)\beta_1^*] = 1.2 + [0.6 * 0 + 0.4 * 0] = 1.2.$$

Finally, we can calculate the topic difficulty d_1 for the first iteration:

$$d_1 = [1 - df(\mathcal{R}_1^*)]\gamma_1^* = [1 - 0.0045] * 1.2 = 1.194.$$

Iteration 2. In the second iteration ($j = 2$), we again first have to build \mathcal{T}_2 and \mathcal{R}_2 by intersecting the set of all relevant images of the most significant topic element of the previous iteration \mathcal{R}_1^* with all the sets of relevant images of the previous iteration $\mathcal{R}_{1,k}$, except $\mathcal{R}_1^* = \mathcal{R}_{1,3}$ itself:

$$\mathcal{T}_2 = \{(guide, photo), (guide, female)\},$$

$$\mathcal{R}_2 = \{\mathcal{R}_{2,1}, \mathcal{R}_{2,2}\} = \{\mathcal{R}_1^* \cap \mathcal{R}_{1,1}, \mathcal{R}_1^* \cap \mathcal{R}_{1,2}\}.$$

We again compare the cardinalities of both sets $\mathcal{R}_{2,1}$ and $\mathcal{R}_{2,2}$: there are 90 images in the *IAPR TC-12 image collection* that show *photos of guides* ($|\mathcal{R}_{2,1}| = 90$) and 26 images that show *female guides* ($|\mathcal{R}_{2,2}| = 26$). The most significant element of the second iteration is therefore *female guides*,

$$\mathcal{R}_2^* = \mathcal{R}_{2,2},$$

with a conditional linear document frequency df of:

$$df(\mathcal{R}_2^*) = \frac{|\mathcal{R}_{2,2}|}{|\mathcal{R}_1^*|} = \frac{|\{\text{images of female guides}\}|}{|\{\text{images of guides}\}|} = \frac{26}{90} = 0.289.$$

A search for *female guides* would return 5 direct and relevant hits from the collection ($|\mathcal{N}_{D_1^*} \cap \mathcal{R}_1^*| = 5$), and there are 26 images that are directly or indirectly annotated with *female guides* ($|\mathcal{N}_1^* \cup \mathcal{R}_1^*| = 26$). Thus, the factor for vocabulary mismatch and incomplete and incorrect annotation is

$$\alpha_2^* = 1 - \frac{|\mathcal{N}_{D_2^*} \cap \mathcal{R}_2^*|}{|\mathcal{R}_2^*|} = 1 - \frac{5}{26} = 1 - 0.19 = 0.81,$$

the factor for word ambiguity

$$\beta_2^* = 1 - \frac{|\mathcal{R}_2^*|}{|\mathcal{N}_2^* \cup \mathcal{R}_2^*|} = 1 - \frac{26}{26} = 1 - 1 = 0,$$

the factor for the annotation gap

$$\gamma_2^* = \eta + [\theta\alpha_2^* + (1 - \theta)\beta_2^*] = 1.2 + [0.6 * 0.81 + 0.4 * 0] = 1.686,$$

and the difficulty for the second iteration

$$d_2 = [1 - df(\mathcal{R}_2^*)] * \gamma_2^* = [1 - 0.289] * 1.686 = 1.199.$$

Iteration 3. In the third iteration ($j = 3$), we again first build \mathcal{T}_3 and \mathcal{R}_3 by intersecting the set of all relevant images of the most significant topic element of the previous iteration \mathcal{R}_2^* with all the sets of relevant images of all elements of the previous iteration $\mathcal{R}_{2,k}$, except $\mathcal{R}_2^* = \mathcal{R}_{2,2}$ itself:

$$\mathcal{T}_3 = \{(guide, female, photo)\},$$

$$\mathcal{R}_3 = \{\mathcal{R}_{3,1}\} = \{\mathcal{R}_2^* \cap \mathcal{R}_{2,1}\}.$$

The *IAPR TC-12 image collection* comprises 26 images that show *photos of female guides* ($|\mathcal{R}_{3,1}| = 26$), and being the only element left, $\mathcal{R}_{3,1}$ is automatically the most significant sentence element for the third iteration as well,

$$\mathcal{R}_3^* = \mathcal{R}_{3,1}$$

showing a conditional linear document frequency of

$$df(\mathcal{R}_3^*) = \frac{|\mathcal{R}_{3,1}|}{|\mathcal{R}_2^*|} = \frac{|\{\text{images of photos of female guides}\}|}{|\{\text{images of female guides}\}|} = \frac{26}{26} = 1.$$

A search for *photos of female guides* would not return any direct and relevant hits from the *IAPR TC-12 image collection* ($|\mathcal{N}_{D_1^*} \cap \mathcal{R}_1^*| = 0$), while it contains 26 images that are directly or indirectly annotated with *photos of female guides* ($|\mathcal{N}_1^* \cup \mathcal{R}_1^*| = 26$). Thus, the factor for vocabulary mismatch and incomplete and incorrect annotation in the third iteration is

$$\alpha_3^* = 1 - \frac{|\mathcal{N}_{D_3^*} \cap \mathcal{R}_3^*|}{|\mathcal{R}_3^*|} = 1 - \frac{0}{26} = 1 - 0 = 1,$$

the factor for word ambiguity

$$\beta_3^* = 1 - \frac{|\mathcal{R}_3^*|}{|\mathcal{N}_3^* \cup \mathcal{R}_3^*|} = 1 - \frac{26}{26} = 1 - 1 = 0,$$

the factor for the annotation gap

$$\gamma_3^* = \eta + [\theta\alpha_3^* + (1 - \theta)\beta_3^*] = 1.2 + [0.6 * 1 + 0.4 * 0] = 1.8,$$

and the difficulty for the third iteration

$$d_3 = [1 - df(\mathcal{R}_3^*)] * \gamma_3^* = [1 - 1] * 1.686 = 0.$$

Final result. The total topic difficulty of topic \mathcal{T} (“photos of female guides”) for the *IAPR TC-12 image collection* \mathcal{I} after three iterations amounts to

$$d(\mathcal{T}, \mathcal{I}) = \sum_{j=1}^3 d_j = 1.194 + 1.199 + 0 = 2.393.$$

5.3 Experimental Validation and Analysis

The validation of the model defined in Section 5.2 comprises two components: first, we report on the level of correlation with system effectiveness in order to indicate the measure’s efficiency at estimating topic difficulty (Section 5.3.1); and second, we compare the results of this model with alternative approaches we attempted as well as with approaches in the existing literature (Section 5.3.2).

In addition, we provide an analysis and propose to classify topics in five different levels of difficulty to illustrate both easy and hard topics (Section 5.3.3); and finally, we point out the benefits and limitations of the novel algorithm (Section 5.3.4).

5.3.1 Correlation with System Effectiveness

The validation in this section is based on the assumption that effective (or good) retrieval by a VIR system is reflected by the value of measures such as MAP , $P(10)$ and $P(20)$. Hence, to validate the proposed measure of topic difficulty, we computed the difficulty of 113 topics from the *ImageCLEF image retrieval benchmark* and correlated these with the results of 132 runs for monolingual English retrieval (we had showed in [149] that topic difficulty is language-dependent).

These results are based on two image collections: we first validated the measure with 53 query topics from the *ImageCLEF 2004 and 2005 ad-hoc retrieval tasks* using the SAC (see Section 3.2.2) as document collection and correlated them with the results of 83 submitted runs for monolingual English retrieval in both years. We then used the novel measure to predict the retrieval effectiveness for additional 60 query topics in the *ImageCLEF 2006 ad-hoc retrieval task (ImageCLEFphoto)* using the *IAPR TC-12 Benchmark* (see Chapter 7), and subsequently validated this prediction with the results of another 49 monolingual English runs in 2006. Figure 5.2 provides a graphical overview of the difficulty of these 113 query topics and the corresponding MAP results of the 132 runs we evaluated.

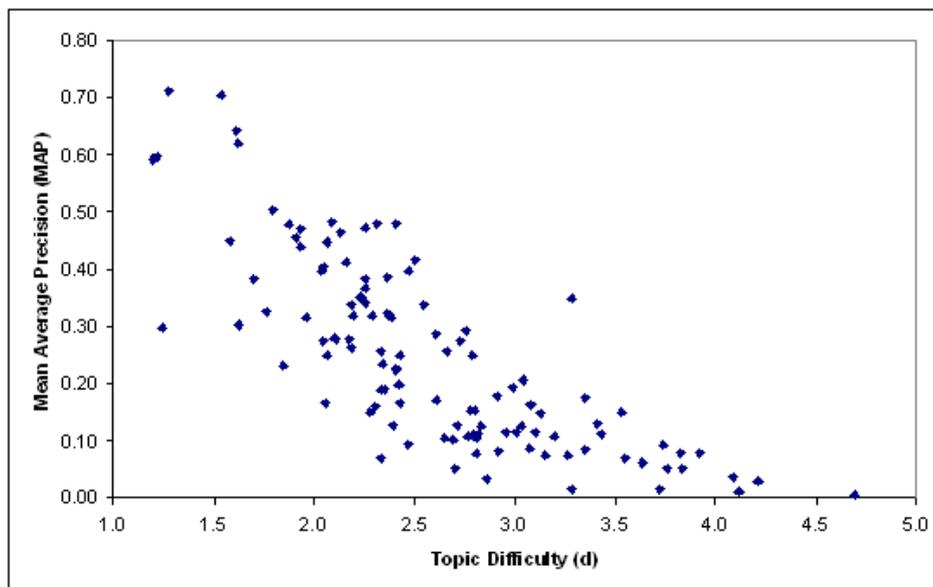


Figure 5.2: MAP and difficulty of 113 ImageCLEF topics (2004-2006).

Finally, we combined the results from all topics and correlated them with their corresponding difficulty values to show that the measure can be used on different image collections. The correlation values $\rho(X, Y)$ were calculated using *Pearson's product moment correlation*, which corresponds to the covariance of the two considered variables X and Y divided by their standard deviations σ_X and σ_Y

$$\rho(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{E((X - \mu_X)(Y - \mu_Y))}{\sigma_X \sigma_Y} \quad (5.17)$$

where $-1 \leq \rho(X, Y) \leq 1$, $E(X) = \mu_X$, $E(Y) = \mu_Y$, and E denotes the expected value for the variable. A strong correlation is indicated by a high absolute value for ρ ; thus, we aim for a strong negative correlation: the higher the topic difficulty measure, the lower the precision values for the results.

Year	#runs	#topics	$\rho(d, MAP)$	$\rho(d, P10)$	$\rho(d, P20)$
2004	30	25	-0.818	-0.692	-0.593
2005	62	28	-0.764	-0.671	-0.700
2006	40	60	-0.711	-0.647	-0.660
Total	132	113	-0.783	-0.686	-0.666

Table 5.1: Correlations for the topic difficulty measure.

Table 5.1 illustrates the correlations of the topic difficulty values and the main performance measures of the *ImageCLEF* 2004–2006 results using two benchmark collections. The strong negative correlation of the proposed measure with the results over a period of three years and using two different collections demonstrates that the proposed algorithm is both robust and applicable across collections. We attribute this to the fact that the measure considers document frequencies as well as the quality of the logical image representations as indicated by the annotation gap factor (γ). All correlations are significant at the 0.01 level, indicating a high level of confidence in the correlation.

The correlation values for $P(10)$ and $P(20)$ are weaker than for MAP scores. This is likely because $P(20)$ can exhibit misleading values if the number of relevant images for a topic is less than 20, and $P(10)$ can be insignificant for queries with too many direct hits.

5.3.2 Alternative Approaches

The algorithm described so far is entirely manual and involves a substantial amount of effort to generate the difficulty scores: performing multiple queries for each topic sentence element and judging as many relevant documents as possible (although judging images for relevance is typically far quicker than text documents, particularly identifying an irrelevant image). The main reason for this has been to establish an upper boundary for correlation with system effectiveness (assuming that automating the algorithm will only cause the correlation to reduce).

To carry out manual searching for computing the difficulty score, we used an ISJ system that ranks images based on the *BM25* weighting operator (see Section 6.3.4 for a detailed description of the system). Given that generating topics for IR benchmarks is by its very nature a manual task, we contrast the proposed approach with a variety of measures based on varying degrees of (perceived) manual effort and group these measures into three classes: low-cost, medium-cost and high-cost.

Measures classified as *low-cost* require minimal manual effort: multiple queries for each topic are not generated, but rather the query is used directly as a whole. Relevance judgements for each topic only need to be performed up to the first 20 ranked images (and captions), displayed on a single results page.

Measures classified as *medium-cost* require more time and effort: individual queries are issued for each sentence element (*i.e.* at each iteration) in the topic and as many relevant images are identified from the results as possible. This category, however, does not require that multiple terms are generated for each topic element.

Measures in the *high-cost* category take the longest to compute with the most manual effort: multiple queries must be issued with each sentence element and, as with the medium-cost approaches, as many relevant images as possible must be found each time.

Low-Cost Approaches

First, we examined several manual prediction methods that can be calculated with a low degree of effort: for each topic, we determined the rank of the first relevant document ($Rank_1$), the rank for the tenth relevant document ($Rank_{10}$), the precision at rank 10 ($P(10)$), the precision at rank 20 ($P(20)$), the total number of documents returned ($UNION$), the number of topic elements (K), and the number of topic terms ($\#words$) based on using ISJ for each query.

Year	$Rank_1$	$Rank_{10}$	$P(10)$	$P(20)$	UNION	K	$\#words$
2006	-0.310	-0.687	-0.522	-0.648	-0.077	-0.430	-0.005
2005	-0.504	-0.577	-0.661	-0.576	-0.292	-0.558	-0.382
2004	-0.458	-0.481	-0.665	-0.565	0.229	-0.525	0.319
AVG	-0.424	-0.582	-0.616	-0.596	-0.047	-0.504	-0.023

Table 5.2: Low-cost approaches and correlations with MAP.

Table 5.2 shows that $Rank_{10}$, $P(10)$ and $P(20)$ have strong correlations of about -0.6 ; $Rank_1$ is not as reliable, while the total number of returned documents for all query terms ($UNION$) only produces random predictions.

It is further noticeable that estimations based on the number of individual query terms ($\#words$) do not show any correlation at all, while the use of the number of grammatical sentence elements (K) exhibits correlation values between -0.43 and -0.56 , which corresponds with previous research such as [116, 149].

Medium-Cost Approaches

Next, the following approaches, requiring slightly more effort than those in Table 5.2, were tested in order to improve the chance of difficulty prediction: first, we calculated the sum of the probabilities of all the query elements (approach P1), and then the sum of $tf-idf$ of all the query elements (IDF1), whereby each query element was treated equally in both approaches.

Both approaches showed only very weak correlations (compare Table 5.3), which was especially surprising for the $tf-idf$ approach, because it was well established in the field of information (text) retrieval. This might be due to the very short image

representations in which the term frequencies tf degenerate to insignificance (in most alphanumeric image representations, $tf = 1$), with only the idf remaining as the discriminating factor².

Year	P1	IDF1	LSE	MSW	LSW	MSE
2006	-0.253	-0.346	-0.516	-0.490	-0.365	-0.647
2005	-0.371	-0.454	-0.504	-0.479	-0.380	-0.715
2004	0.017	-0.078	-0.492	-0.468	-0.395	-0.660
AVG	-0.202	-0.293	-0.504	-0.479	-0.380	-0.674

Table 5.3: Medium-cost approaches and correlations with MAP.

We hence discarded the approaches that would treat each sentence element equally and replaced them with an element priority scheme. We consequently based the next approaches on the probabilities of the least (LSW) or most (MSW) significant topic term as well as on the least (LSE) or most (MSE) significant topic elements respectively, and performed the algorithm as described in Section 5.2.3, only without the integration of the annotation gap. We also tried the same four approaches using $td-idf$ weights instead of probabilities, but they again showed a much weaker or no correlation (and are therefore not included in the table).

The approach based on the probabilities of the *most significant element* (MSE) showed the most promising results with a negative correlation of nearly -0.68 , an improvement of around 17% in comparison with the low-effort approaches.

High-Cost Approaches

Due to the results of the medium-cost approaches, only the approaches based on the most significant element were considered for the high-cost approaches, and further effort was undertaken to improve the correlation results (*e.g.* incorporation of the annotation gap γ): using the algorithm as described in Section 5.2.3 with the probabilities (P2) or $tf-idf$ (IDF2) multiplied with average γ for individual elements, the probabilities (P3) or $tf-idf$ (IDF3) multiplied with γ at iterations, the sum of probabilities and γ (P4), the sum of $td-idf$ and γ (IDF4), and the probabilities (P5) or $tf-idf$ (IDF5) multiplied with γ for individual elements.

²We therefore only use IDF to denote the approaches using $tf-idf$ weighting.

Year	P2	IDF2	P3	IDF3	P4	IDF4	P5	IDF5
2006	-0.685	-0.181	-0.711	-0.255	-0.546	-0.341	-0.701	-0.252
2005	-0.752	-0.362	-0.764	-0.290	-0.653	-0.222	-0.763	-0.413
2004	-0.674	-0.110	-0.818	-0.186	-0.759	-0.345	-0.735	-0.192
AVG	-0.704	-0.218	-0.783	-0.244	-0.653	-0.303	-0.733	-0.286

Table 5.4: High-cost approaches and correlations with MAP.

Table 5.4 shows the correlations with the average MAP, with the algorithm P3 as described in Section 5.2 exhibiting a negative correlation of almost -0.8 , a further substantial improvement in comparison to the low and medium cost approaches.

Finally, using *idf/tf* approaches, again, showed considerably lower correlations than approaches based on simple probabilities.

5.3.3 Topic Difficulty Analysis

This section utilises the previous results to categorise topics into classes of difficulty to illustrate both easy and hard topics. Such a classification is useful insofar as absolute values do not directly indicate the difficulty of a topic.

d	Level	MAP	P(10)	P(20)
$0 \leq d < 1$	very easy	N/A	N/A	N/A
$1 \leq d < 2$	easy	0.473 (0.15)	0.542 (0.12)	0.480 (0.12)
$2 \leq d < 3$	medium	0.248 (0.12)	0.354 (0.20)	0.319 (0.18)
$3 \leq d < 4$	hard	0.109 (0.07)	0.178 (0.11)	0.166 (0.10)
$d \geq 4$	very hard	0.020 (0.01)	0.065 (0.06)	0.060 (0.05)

Table 5.5: Topic difficulty levels.

Table 5.5 shows the classification of topics in five different difficulty levels, together with their average results from *ImageCLEF* per category (and the standard deviation in parenthesis). The results of each of the categories correlate well with *MAP*, *P(10)* and *P(20)* for both the *SAC* of historic photographs and the *IAPR TC-12* collection of generic photographs.

Easy Topics

In general, easy topics are those in which the majority of (or all) the query elements are direct hits in a well annotated database. Consider the examples for easy topics

Topic	d	MAP	P(10)	P(20)
photos of radio telescopes	1.199	0.5914	0.5914	0.5350
animal statue	1.222	0.5974	0.7339	0.7113
Rome in April 1908	1.698	0.3840	0.4167	0.3583

Table 5.6: Examples of easy topics.

given in Table 5.6: all three topics have a very simple grammatical structure, and all their sentence elements (animal statue, Rome, April, 1908, radio telescope) happen to be direct hits and can therefore easily be retrieved using a direct keyword search without any further sophisticated technology. Topic difficulty values are especially low when the most significant element does not occur often and implicitly forms the set of relevant images for that topic (*e.g.* all the images showing a radio telescope are relevant).

Difficult Topics

Difficult topics as shown in Table 5.7 often exhibit a more complex sentence structure and require more sophisticated approaches to accurately retrieve relevant images. For instance, spatial relations using non-containment operators (such as near,

Topic	d	MAP	P(10)	P(20)
tourist accommodation near Lake Titicaca	4.696	0.0055	0.0100	0.0075
people in bad weather	4.213	0.0279	0.0875	0.0813
royal visits to Scotland (except Fife)	3.756	0.0503	0.0129	0.0113
church with more than two towers	3.739	0.0912	0.0900	0.0888

Table 5.7: Examples of difficult topics.

around, north of, *etc.*) make it almost impossible to retrieve relevant images unless the retrieval system exhibits spatial awareness (*e.g.* for the query “tourist accommodation *near* Lake Titicaca”).

Then, IR systems are not able to handle negators (*e.g.* not, except) by default, because to do so reliably is well beyond the scope of current NLP, and negators are thus treated as standard query terms; consequently, many irrelevant images might be retrieved. For example, the topic “royal visits to Scotland (*except* Fife)” showed very low precision results as many highly ranked images were, indeed, from Fife.

Numerals denoted as expressions (“church with *more than two* towers”), high-level semantic concepts that are not captured in the logical image representations (“*bad* weather”, “*creative* pictures”, *etc.*) or vocabulary mismatches between topics and representations also appear to cause problems for concept-based image retrieval systems.

Finally, a small target set of relevant images does not necessarily make a topic easier because the likelihood of finding a correct image is reduced.

5.3.4 Benefits and Limitations

The proposed method displays a strong negative correlation between system effectiveness (as quantified using $P(10)$, $P(20)$ and MAP) and topic difficulty, giving an upper boundary of the correlation which can be achieved using a costly manual approach. Having such an accurate measure enables the creators of IR benchmarks to carefully select topics, especially for concept-based image retrieval.

One drawback of the algorithm can be seen in the amount of work that goes into determining the difficulty for a single topic as the collection frequencies and the number of direct hits must be calculated for each grammatical element of the topic. While the textual part can certainly be automated, identifying relevant images does, at this stage of research, still involve human interaction. Although this approach is well-suited to the already manual task of topic selection in benchmarking, it is not suitable for real-time computation of topic difficulty.

By comparing the novel algorithm with alternative approaches of varying levels of manual effort (or cost) associated with them, methods which involve less manual effort are certainly successful, but at a cost of lowering correlation and ultimately being less successful at predicting system effectiveness.

5.4 Topic Design Methodology

The goal of the topic creation process was to provide a number of representative search topics for the photographic ad-hoc retrieval task of *ImageCLEF* 2006. This task (called *ImageCLEFphoto*) is 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 the search topics are not known to the system in advance. The specific goal of the simulation is: given an alphanumeric statement (and/or sample images) describing a user information need, find as many relevant images as possible from a generic photographic collection (with the query language either being identical or different from that used to describe the images) [60].

This scenario closely corresponds to that of customers and employees of *viventura* requesting images from the *IAPR TC-12 photographic collection*. Since there is no prior work to report on search behaviour for this particular scenario, we first set up a logging function to monitor the user information needs and to further create a pool of potential topic candidates (Section 5.4.1), before we developed 60 representative search topics against a number of dimensions (Section 5.4.2) which also included the novel topic difficulty measure.

5.4.1 User Need Analysis

This analysis originates from a web-based interface to the *IAPR TC-12 image collection* which was used by employees and customers of *viventura*. The data was collected from 1 February to 15 April 2006, with the log file containing 980 unique queries. Most of the queries were performed in German or Spanish and later translated to English. The average query length for English was 2.45 words, with a standard deviation of 1.61 words; the longest query comprised 12 words and the shortest was one word.

Search Characteristics

According to the log file, the following search characteristics could be identified for retrieval from the *IAPR TC-12 photographic collection*:

- *Query types*: most of the queries are short noun phrases, often with a place adjunct.
- *Level*: queries are exclusively on (*pre-*)*iconographic* image level (compare Section 2.2.3): only general and specific requests are made, but none for images with emotional or symbolic significance (*iconological* level).
- *Length*: the majority of English queries (59%) is between 2 and 5 words. 37% are single word queries, the rest (4%) is 6 words or longer. The German queries are slightly shorter.
- *Nouns*: people search for both general nouns and proper names.
- *Adjectives*: only a few queries use adjectives, mostly with colour information. Adjectives are mainly used in queries including solely one or two objects, but not in longer ones.
- *Verbs*: only a few queries use verbs to indicate an action.
- *Geographic constraints*: many queries involve additional geographic information, some with specific descriptions (“*in* La Paz”), while others make use of spatial operators (“*near* Lake Titicaca”, “*around* Quito”).
- *Prepositions*: used irregularly, some people make use of them (“churches *in* Ecuador”), others do not (“churches Ecuador”).
- *Time constraints*: people generally do not (yet) look for pictures restricted to certain periods or years.

Search Patterns

The main search requests are for general and specific tourist destinations, people, landscapes, regions, accommodations, animals, social projects, actions, and specific objects as well as abstract terms. The majority of requests in one of these search areas thereby follows a specific pattern as illustrated in Table 5.8.

Search Pattern	Example
LOCATION	Rio de Janeiro
COUNTRY	Brazil
REGION	Patagonia
LOCATION - COUNTRY	Rio de Janeiro, Brazil
TOURIST DESTINATION	Mitad del mundo
TOURIST DESTINATION - LOCATION	Mitad del mundo, Quito
TOURIST DESTINATION - COUNTRY	Mitad del mundo, Ecuador
ACCOMMODATION	Host families
ACCOMMODATION - SPECIFICATION	Host families with swimming pool
ACCOMMODATION - LOCATION	Host families near Lake Titicaca
ANIMAL	Boobies
ANIMAL - LOCATION	Boobies in Ecuador
ANIMAL - SPECIFICATION	Blue-footed boobies
ANIMAL - SPECIFICATION - LOCATION	Blue-footed boobies in Ecuador
PEOPLE	Surf instructor
PEOPLE - SPECIFICATION	Godchildren with red cap
PEOPLE - LOCATION	Godchildren in Peru
PEOPLE - PROPER NAMES	André Kowitz
PEOPLE - PROPER NAMES - LOCATION	André Kowitz in Botogá
OBJECT	Church
OBJECT - SPECIFICATION	Church with one tower
OBJECT - LOCATION	Church in Ecuador
ACTION	Surfing
ACTION - LOCATION	Surfing in Brazil
SOCIAL PROJECT	Kindergarten project
SOCIAL PROJECT - LOCATION	Kindergarten project in Quito
ABSTRACT TERM	Football
ABSTRACT TERM - LOCATION	Football in Ecuador
LANDSCAPE	Mountain scenery
LANDSCAPE - LOCATION	Mountain scenery in Patagonia

Table 5.8: Search patterns.

It can, again, be noticed that many search patterns exhibit some kind of geographic constraint, which concurs with previous studies for retrieval from generic photographic collections [379, 503].

5.4.2 Topic Development and Dimensions

The log file analysis did not only offer direct insight into search patterns and characteristics specific to the *IAPR TC-12 image collection*, it also provided a pool of 980 topic candidates which formed the foundation for the topic development process. To provide an element of control over the selection from these topics candidates,

ID	Topic Title	ID	Topic Title
1	accommodation with swimming pool	31	volcanos around Quito
2	church with more than two towers	32	photos of female guides
3	religious statue in the foreground	33	people on surfboards
4	group standing in front of mountain landscape in Patagonia	34	group pictures on a beach
5	animal swimming	35	bird flying
6	straight road in the USA	36	photos with Machu Picchu in the background
7	group standing in salt pan	37	sights along the Inka-Trail
8	host families posing for a photo	38	Machu Picchu and Huayna Picchu in bad weather
9	tourist accommodation near Lake Titicaca	39	people in bad weather
10	destinations in Venezuela	40	tourist destinations in bad weather
11	black and white photos of Russia	41	winter landscape in South America
12	people observing football match	42	pictures taken on Ayers Rock
13	exterior view of school building	43	sunset over water
14	scenes of footballers in action	44	mountains on mainland Australia
15	night shots of cathedrals	45	South American meat dishes
16	people in San Francisco	46	Asian women and/or girls
17	lighthouses at the sea	47	photos of heavy traffic in Asia
18	sport stadium outside Australia	48	vehicle in South Korea
19	exterior view of sport stadia	49	images of typical Australian animals
20	close-up photograph of an animal	50	indoor photos of churches or cathedrals
21	accommodation provided by host families	50	cathedrals
22	tennis player during rally	51	photos of goddaughters from Brazil
23	sport photos from California	52	sports people with prizes
24	snowcapped buildings in Europe	53	views of walls with unsymmetric stones
25	people with a flag	54	famous television (and telecommunication) towers
26	godson with baseball cap	55	drawings in Peruvian deserts
27	motorcyclists racing at the Australian Motorcycle Grand Prix	56	photos of oxidised vehicles
28	cathedrals in Ecuador	57	photos of radio telescopes
29	views of Sydney's world-famous landmarks	58	seals near water
30	room with more than two beds	59	creative group pictures in Uyuni
		60	salt heaps in salt pan

Table 5.9: ImageCLEFphoto 2006 topics.

we identified and considered the following dimensions for the selection of the final set of topics (see Table 5.9) that was eventually distributed to the participants: the

topic origin, geographical constraints, the “visuality” of the topic, the estimated number of relevant images, the degree of representation “completeness”, additional text retrieval challenges, the difficulty of the topic, feedback from previously held evaluations, and, last but not least, past research on image retrieval search such as [104]. The exact distribution of the topics over these dimensions (together with the corresponding results for each of these dimensions) can be found in Appendix A. Most of the following is taken from [61].

Topic Numbers and Origin

As for the number of the final topic set, we decided to select 60 topics to represent typical search requests for the *IAPR TC-12 Benchmark*. This number is slightly higher than the preferred default (*i.e.* 50 topics) by TREC [475] in order to further increase the reliability of our results.

To make the task realistic, we took 40 topics directly from the log file (semantically equivalent but perhaps with slight syntactic modification, *e.g.* “lighthouse sea” to “lighthouses at the sea”) and derived 10 further topics from entries in the log file (*e.g.* “straight roads in Argentina” changed to “straight roads in the USA”). The remaining 10 topics were not taken directly from the log file, but based on domain knowledge of the topic authors and created to test various aspects of text and image retrieval (*e.g.* “black and white photos of Russia”).

Text Retrieval Challenges

For many of the topics, successful retrieval using concept-based IR methods will require the use of query analysis (*e.g.* expansion of query terms or logical inference). These reflect examples found in the log files, *e.g.* for the query “group pictures on a beach”, many of the alphanumeric image representations will not contain the term “group” but rather terms such as “men” and “women” or the names of individuals.

Similarly for the query “accommodation with swimming pool” (also from the log file), the query will result in limited effectiveness unless “accommodation” is expanded to terms such as “hotel” or “hostel”. Queries such as “images of typical

Australian animals” require a higher level of inference and access to world knowledge (this query is not found in the log file but could be a feasible request by users of an image retrieval system).

Apart from the aforementioned investigation of general versus specific concepts and the additional challenge of vocabulary mismatches between query topics and logical image representations, we also offered various other challenges for concept-based image retrieval such as the inclusion of ambiguous terms like “San Francisco” in the topic ”people in San Francisco” (which can either refer to the Californian city but also to South American churches consecrated to Francis of Assisi) and the use of abbreviations such as “USA” in the topic “straight roads in the USA”.

Further multilingual aspects that we considered for the translation of topics include: dealing with proper names, compound words, morphological variants, idioms, acronyms and equivalent syntactic and semantic expressions.

Visual Retrieval Challenges

We also classified all topics regarding how “visual” they were considered to be. An average rating between 1 and 5 was obtained, which we based on the retrieval score from a baseline CBIRS (FIRE, see Section 2.7.5) and on the opinion of three experts in the field of image analysis, who we had asked to rate these topics according to the following scheme: CBIR would produce

- (1) very bad or random results,
- (2) bad results,
- (3) average results,
- (4) good results,
- (5) very good results.

Based on these findings, we then classified a total of 30 topics as “semantic” (levels 1 and 2) for which visual approaches would be highly unlikely to improve results

(*e.g.* “cathedrals in Ecuador”), 20 topics as “neutral” (level 3) for which visual approaches may or may not improve results (*e.g.* “group pictures on a beach”), and 10 topics as “visual” for which content-based approaches would be most likely to improve retrieval results (*e.g.* “sunset over water”).

Geographic Constraints

Similar to previous analyses of search log files (see also Section 3.3.2), we found many search requests to exhibit some kind of a geographical constraint (*e.g.* specifying a location).

Therefore, we selected 24 topics with a geographic constraint (*e.g.* “tourist accommodation *near Lake Titicaca*” specifies a location and spatial operator *near*), 20 topics with a geographic feature or a permanent man-made object (*e.g.* “group standing in *salt pan*”) and 16 topics with no geography (*e.g.* “photos of female guides”).

Topic Difficulty

Then we examined the difficulty of the topics and categorised them with respect to the novel measure defined in Section 5.2: 4 topics were classified as “easy” (*e.g.* “bird flying”), 21 as “medium” (*e.g.* “pictures taken on Ayers Rock”), 31 as “hard” (*e.g.* “winter landscape in South America”) and 4 as “very hard” (*e.g.* “tourist accommodation near Lake Titicaca”). See Table 5.5 in Section 5.3.3 for the exact definition of these topic difficulty levels.

Representation Completeness

Another dimension considered was the distribution of the topics as regards the level of representation “completeness” of relevant images (see Section 6.1.2) for the particular topics. We introduced this dimension to be able to observe whether more visual approaches would improve the retrieval results for topics that predominately target images with incomplete semantic representations.

Hence, we provided 18 topics in which all relevant images had complete representations, 10 topics with 80% - 100% of the relevant images having complete representations, a further 19 topics with 60% - 80% of the relevant images with complete representations, and 13 topics with less than 60% of the relevant images with complete representations.

Size of Target Set

The estimated number of relevant images for each topic (*i.e.* the target set size) is a dimension which we primarily considered for organisational purposes: we aimed for a target set size between 20 and 100 relevant images and thus had to further modify some of the topics (broadening or narrowing the concepts). The minimum was chosen in order to be able to use $P(20)$ as a performance measure, whereas the upper limit of relevant images should limit the retrieval of relevant images by chance and to keep the relevance judgment pools to a manageable size.

Participant Feedback

Participants had suggested in prior events that we provided groups of similar topics in order to facilitate the analysis of weakly performing queries. We also considered this input in the topic development process and clustered the topics in groups of up to five topics. An example for topics in one cluster is: “people in bad weather”, “destinations in bad weather”, “Machu Picchu in bad weather”.

5.5 Summary

In this chapter, we have presented a model that we established in order to facilitate the *topic creation process* for image retrieval evaluation events; this comprised the identification of several query dimensions as well as the analysis of a log file to base the topic creation process on realistic user information needs for retrieval from the *IAPR TC-12 image collection*. Taking these potential topics from the log file into consideration, we then created a set of representative query topics against the query dimensions we had identified before.

The largest contribution of this chapter, however, is the definition of a novel measure to quantify topic difficulty for TBIR based on both linguistic features of the topic and statistical information gained from the corresponding document collection. The novel measure displays a strong negative correlation between topic difficulty and system effectiveness as quantified using MAP , $P(10)$ and $P(20)$, and gives an upper boundary of the correlation which can be achieved using a costly manual approach. The difficulty of concept-based image retrieval had not been studied to date, and we argue that having such an accurate measure enables the creators of concept-based image retrieval evaluation events to carefully select topics, making topic difficulty one of the most significant dimensions in the topic creation process.

The development of the *IAPR TC-12 Image Benchmark*, including a freely available image collection together with a set of representative query topics, has certainly brought a massive contribution to the field of VIR. However, the creation of these major benchmark components would have not been possible without the use of a custom-built parametric benchmark administration system, which is further described in the next chapter.

Chapter 6

Parametric Benchmark Design and Architecture

The two previous chapters have reported on our methodology that (1) enabled the careful and consistent design and development of a representative document collection (*i.e.* images and their semantic descriptions) to allow for the evaluation of VIR from generic photographic collections (Chapter 4), and (2) facilitated the creation of a natural, balanced topic set accurately reflecting real world user statements of information needs (Chapter 5).

However, one vital aspect we have not dealt with thus far is the underlying technology that made the realisation of the aforementioned methodology possible: a parametric benchmark administration system we specifically designed and implemented in order to

- support the initial incremental development of the benchmark;
- facilitate and guide the ongoing management of the major benchmark components;
- enable a deeper understanding of the complex processes associated with the evaluation of VIRS;
- allow for the dynamic reaction to changed evaluation requirements.

Hence, in this chapter, we will introduce the novel architecture of a parametric benchmark system. This first comprises the identification of the fundamental bench-

mark parameters (Section 6.1) and their representation in a relational database (Section 6.2). Based on this underlying relational architecture, we then present an overview of the functionality of the benchmark administration system (Section 6.3), and we finally point out the benefits of parametric image benchmarks (Section 6.4).

6.1 Benchmark Parameters

The benchmark management and administration system presented in Section 6.3 currently supports the specification of several parameters, which can be used

- to create different subsets of the image collection (Section 6.1.1),
- to facilitate the variation of logical image representations (Section 6.1.2), and
- to develop and analyse the representative query topics (Section 6.1.3).

This section briefly introduces these parameters and points out the existing (or potential) use of the corresponding subsets created.

6.1.1 Collection Parameters

The benchmark administration system allows the generation of image subsets of the *IAPR TC-12 photographic collection* with respect to the following parameters.

Collection Size

The size of the image collection might constitute the most obvious parameter: theoretically, any size between zero and the total number of images in the collection ($N = 20,000$) could be selected, although one should opt for at least a minimum of 1,000 images in order to comply with the original benchmark requirements (see Section 4.1.2).

Image subsets can either be specifically selected by their position in the database (*e.g.* the first 5,000 images) or by their unique identifiers (*e.g.* images having a unique identifier between 7500 and 12500), or they can be randomly selected (*e.g.* 5,000 randomly selected images from the collection).

In most cases, however, the image contents, rather than the collection size, is the primary reason for the selection of a subset.

Image Category

One example for such an image content parameter that allows for the creation of a collection subset is the *image category*. The choice of this parameter can inevitably create a very domain-specific subset from the rather generic document collection (for example, by only selecting animal photos). Such subsets can potentially provide a useful resource for very specific evaluation goals (such as animal recognition, see [160]).

Image Complexity

Another parameter that can be used to create various subsets, image complexity, is based on the number of objects and relationships illustrated in the images.

For example, only images which actually contain at least one relationship between objects ($N_R > 0$) can be considered in a subset for the evaluation of images with complex image contents, while a subset with images only containing one object ($N_O = 1$) could be created for and used in current object recognition or automatic annotation tasks.

Location

Subsets according to geographic locations can be generated for researchers who are only interested in retrieval of images from a particular location (*e.g.* South America), region (*e.g.* Patagonia) or country (*e.g.* Argentina). Such subsets could, for instance, be interesting for *geographic information systems* (GIS).

Time

The time parameter can be used for subset generation to carry out an evaluation of the retrieval of images that originate from different years or decades: for example, the evaluation of retrieval from images from just 2000 and 2005 in order to investi-

gate whether the change of technology (from analogue to digital cameras) has had an influence on retrieval results.

Subset Combination

The system allows the generation of any combination, intersection or union, of two or more of the aforementioned subsets as well.

6.1.2 Representation Parameters

For each of the images within the *IAPR TC-12 photographic collection*, the benchmark administration system allows the export of the corresponding logical image representation (also called image *caption* or *annotation*) stored in the database to a plain text file with respect to the following parameters.

Representation Type

The most obvious of these parameters is the type of the logical image representation. Currently supported types are free text representations and semi-structured representations (see Section 7.2.2 for examples), with structured representations as defined in MPEG-7, for example, currently being implemented.

Representation Format

Another key parameter for the export of the semantic image descriptions in the database is the representation format (*i.e.* the tags used in semi-structured representations).

This parameter is essential because, should the format requirements for the semantic image descriptions change, only the required representation format settings would need to be readjusted and the corresponding text files could automatically be re-generated with respect to these new settings, without having to access the text files directly.

Representation Language

In a multilingual evaluation environment such as *ImageCLEF*, it is crucial to provide a parameter for the specification of the *language* used for the semantic descriptions of an image. Hence, the current version of the annotation generator also supports the specification of the following representation languages: English, German and Spanish. We used this parameter for *ImageCLEFphoto 2006* (see Section 7.2.2 for examples) and provided the participants with a subset of English and German representations (see Section 7.2).

It is also possible to export the semantic image descriptions to text files whereby the representation language is randomly selected for each individual image. We are planning to use such a subset for *ImageCLEFphoto 2007*.

Representation Completeness

Not all the images in the real-world are perfectly annotated. Thus, in order to provide a more realistic set of data, it is possible to create subsets withholding information regarding the title, the semantic description, additional notes, the date and the location of capturing the image.

We made use of this parameter for *ImageCLEFphoto 2006* and created a set of representation files with varying completeness levels (see Section 7.2). For *ImageCLEFphoto 2007*, we are planning to generate a test collection with only lightly annotated images (only title, notes, location and date fields) to create a slightly different task to those we offered in previous years.

Representation Quality

The semantic image descriptions can also be exported to representative text files with respect to various levels of orthography to examine the ability of retrieval algorithms to deal with typographical errors. The current implementation permits the random injection (addition), deletion and swapping of characters to simulate potential spelling mistakes.

Parameter Combination

The system allows the specification of any combination of two or more of the aforementioned parameters as well.

6.1.3 Query Parameters

The benchmark administration system also allows the specification of parameters to facilitate the creation and analysis of representative query topics. Examples include the origin, geographical constraints, the “visuality”, the estimated number of relevant images, the degree of representation “completeness” and the estimated retrieval difficulty of the topic as well as additional text-retrieval and translation challenges. These parameters have already been discussed in Section 5.4.2.

6.2 System Architecture

The specification of parameters for the creation of subsets in a test collection is only possible if all the relevant information is kept in a dynamic and central environment that allows for the subsequent and automatic generation of the required collection subsets (images and corresponding semantic descriptions) as well as the development and analysis of representative search topics. As a consequence, we made use of a MySQL database¹ to provide such functionality and to facilitate the parameterisation of our test collection.

6.2.1 Collection Management

Figure 6.1 illustrates the physical database model that supports the operation of the collection management module² to facilitate the administration of the collection images and their corresponding semantic representations. Each table has been carefully designed in accordance with the benchmark requirements as well as with the image selection and annotation rules. The highly flexible architecture, which

¹Version 4.1.20.

²The primary keys of the tables are in bold letters, with the lines indicating the relationships between them.

6.2.2 Topic Management

Figure 6.2 illustrates the physical database model that supports the operation of the topic creation and administration module. Each table has been carefully designed

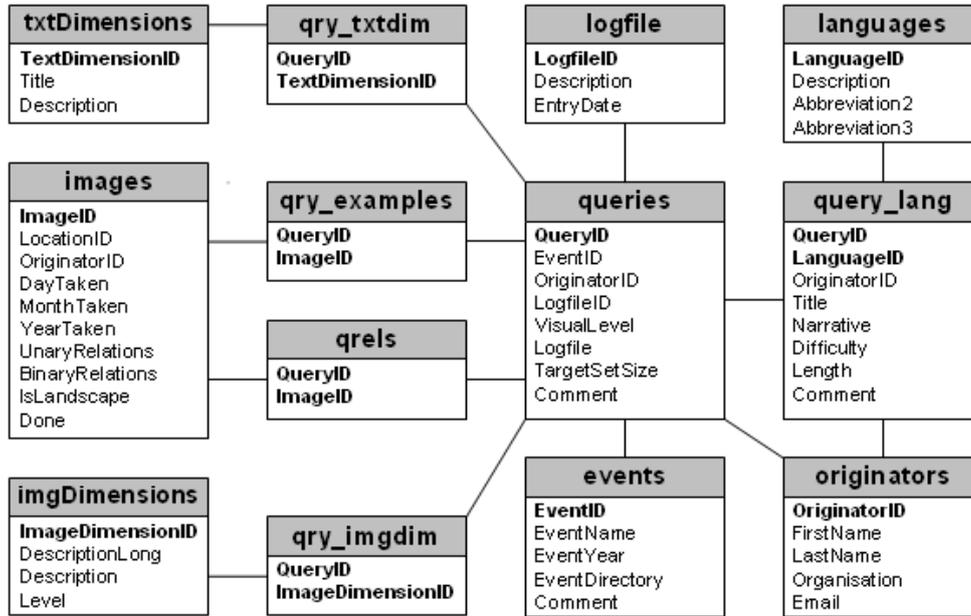


Figure 6.2: Topic management architecture.

to support the creation and translation of topics as well as their analysis based on several dimensions. The highly flexible architecture, again, provides the possibility to include further languages as well as additional dimensions for the analysis of image and text retrieval.

The core of the design evolves around the tables *queries* and *query_lang*, which contain the key information for each of the topics and language specific data for each of their translations. Most of the other entities (that provide further information with respect to events, originators, log file involvement and text and image dimensions) are outsourced for normalisation purposes to guarantee the high flexibility and extensibility of the architecture.

In addition, there are two relationships between the *images* and *queries* tables: *qry_examples* assigns sample images to the topics, while *qrels* holds the information of the estimated set of relevant images for each topic (predefined ground-truth).

6.3 System Overview

This section provides a comprehensive description of the functionality of the benchmark administration and management system which we implemented based on the parametric benchmark architecture presented in Section 6.2.

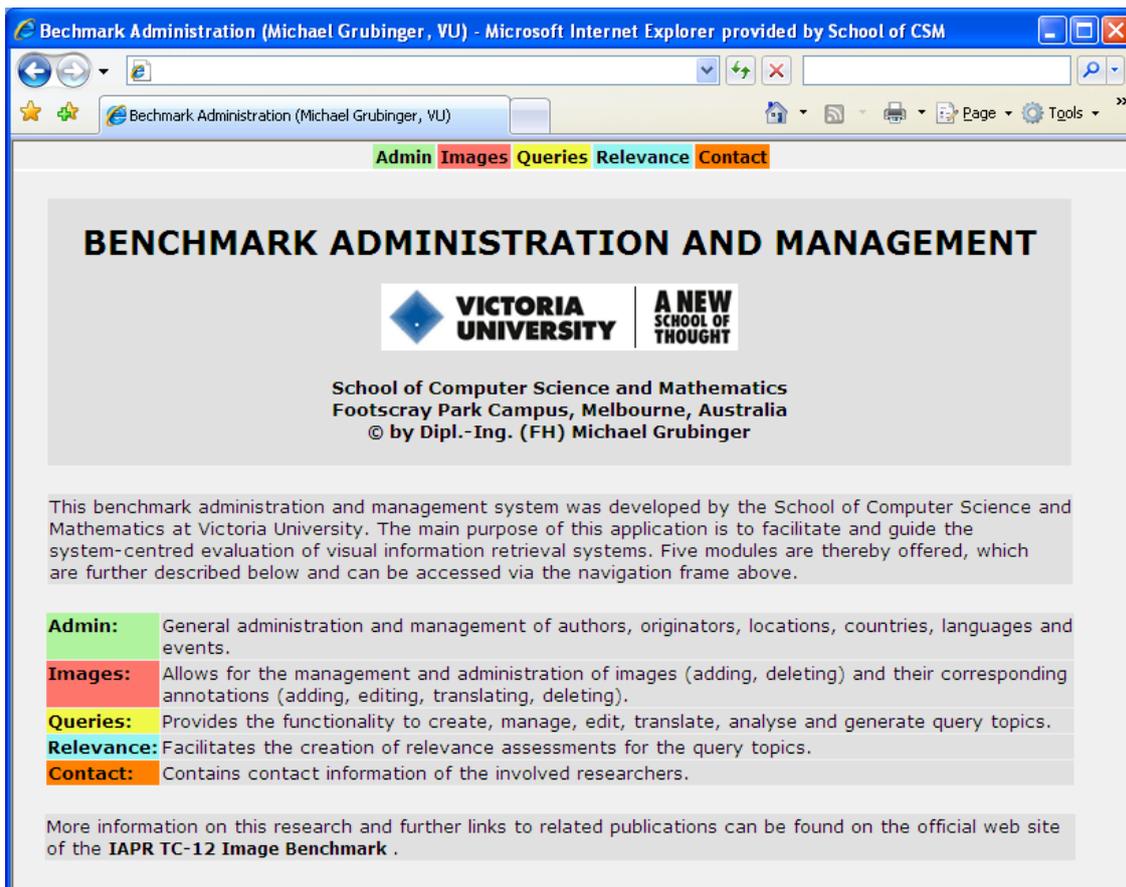


Figure 6.3: Benchmark administration and management.

The main purpose of this application is to facilitate and guide the system-centred evaluation of VIR from generic photographic collections. Figure 6.3 displays the main page of the system, which allows the selection of the following five modules:

- the *admin* module for the management of administrative data (Section 6.3.1);
- the *images* module for the management and administration of the images and their corresponding semantic descriptions (Section 6.3.2);

- the *queries* module for the creation, management and translation of query topics (Section 6.3.3);
- the *relevance* module for the creation of relevance assessments (Section 6.3.4);
- the *contact* module for further information (Section 6.3.5).

The key functionality of each of these modules will be explained in the following sections. Unless indicated otherwise, we used a MySQL database³ to store the underlying information, PHP⁴ for the implementation of the web-based user interface and a Linux server⁵ to host the system files.

6.3.1 General System Administration

This module facilitates the administration of general data required in the remaining modules mentioned below. In particular, it provides the functionality to add, edit and delete the following information:

- **Authors:** the person creating or translating the logical image representations or the query topics;
- **Originators:** the person or entity owning the original copyright of the images;
- **Locations:** the place where the images were taken;
- **Countries:** the country of the locations;
- **Languages:** the language of the logical image representations and the topics;
- **Events:** the evaluation event in which the topics are used.

The basic functionality to add, edit and delete records is very similar for each of these submodules. Rather than listing all of them, we will present one example for the administration of authors.

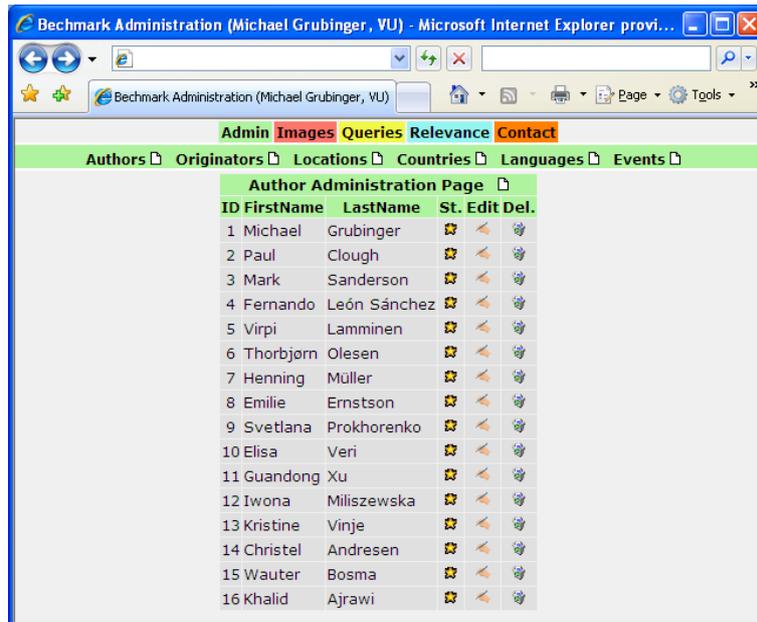
³Version 4.1.20.

⁴PHP: Hypertext Preprocessor, Version 4.3.10-16.

⁵Version 2.6.8-2-686-smp, using Apache as Server API.

Author Overview

If the submodule *authors* is selected in the menu bar on top of the screen, the system responds by displaying an overview of all authors in the database (see Figure 6.4). By default, these records are sorted by their unique identifier (ID), but they



ID	FirstName	LastName	St.	Edit	Del.
1	Michael	Grubinger	*	<	⛕
2	Paul	Clough	*	<	⛕
3	Mark	Sanderson	*	<	⛕
4	Fernando	León Sánchez	*	<	⛕
5	Virpi	Lamminen	*	<	⛕
6	Thorbjorn	Olesen	*	<	⛕
7	Henning	Müller	*	<	⛕
8	Emilie	Ernstson	*	<	⛕
9	Svetlana	Prokhorenko	*	<	⛕
10	Elisa	Veri	*	<	⛕
11	Guandong	Xu	*	<	⛕
12	Iwona	Miliszewska	*	<	⛕
13	Kristine	Vinje	*	<	⛕
14	Christel	Andresen	*	<	⛕
15	Wauter	Bosma	*	<	⛕
16	Khalid	Ajrawi	*	<	⛕

Figure 6.4: Author overview.

can easily be re-sorted by clicking on the column headings (*e.g.* clicking on “First-Name” would sort the records by their first name). In any overview page within the benchmark administration system, the last three columns of each record show the following three clickable symbols: (1) a *yellow star* to display more information of that particular record, (2) a *hand holding a pencil* to edit that record, and (3) a *garbage can* to permanently delete the record.

Delete Authors

If the *garbage can* symbol next to a record is clicked, that record is permanently deleted from the database and the system will subsequently respond with the overview page again (without the deleted record).

Edit Authors

If the *hand symbol* next to a record is clicked, the system responds with a form that allows the user to edit that particular record. In the case of authors, this form is rather simple and allows the specification of the author's first name, last name and email address (see Figure 6.5). One can (1) cancel this action by clicking on the

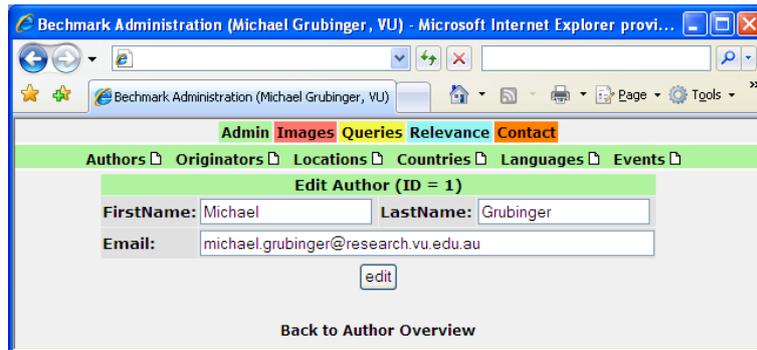


Figure 6.5: Edit authors.

link “Back to Author Overview” or (2) commit the changes by clicking the “edit” button. In both cases, the system will return to the author overview page.

Add Authors

If the *white document* symbol is clicked (either the one next to “Authors” in the submenu or the one next to the table heading “Author Administration Page”), the system responds with an empty form to add a new author (see Figure 6.6). Again,

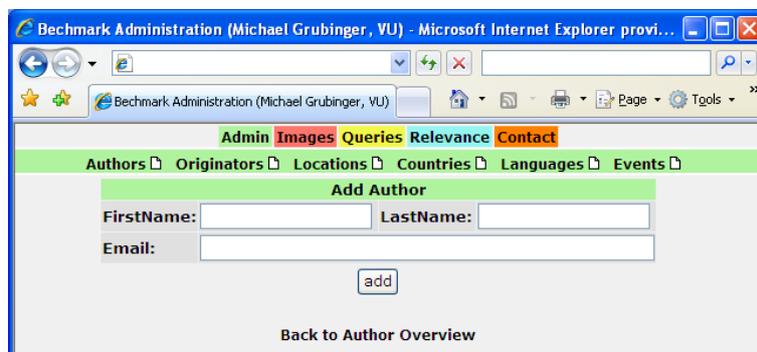


Figure 6.6: Add authors.

this action can either be cancelled by clicking on the “Back to Author Overview”

link or be committed by clicking the “add” button. In both cases, the system will return to the author overview page.

6.3.2 Collection Management and Administration

This section presents one of the key modules of the benchmark administration system: parametric collection management and administration. Not only did this module potentiate the incremental development and facilitate the ongoing maintenance and administration of the document collection (*i.e.* images and their corresponding semantic descriptions), it also made it possible to fulfil the requirements for a parametric benchmark architecture: it allows the specification of a number of parameters that may be adjusted according to different requirements or changing evaluation goals (see also Section 4.1.2).

Image Management

Similar to the general administration module, if the “Images” option is chosen in the main menu, the collection management module displays several submenus and shows an image overview page (see Figure 6.7) by default. The image overview displays 10 clickable image thumbnails in one row and can accommodate up to 100 images per page. The symbols for editing and deleting as well as a status indicator are located below each of these thumbnails.

When an image is added to the collection, a thumbnail is automatically created and both image and thumbnail files are uploaded to the benchmark server, while statistical functions provide additional feedback on whether the insertion of that image complies with the original benchmark specification. Without the use of this module, the incremental development and extension of this document collection would have not been possible without compromising its quality or consistency, because no control over the image selection and annotation rules (see Sections 4.2.3 and 4.3.2) could have been provided. While small collections could certainly be administered by hand (*e.g.* by manually adding new images), the systematic and controlled insertion offered by this collection management system gains more and

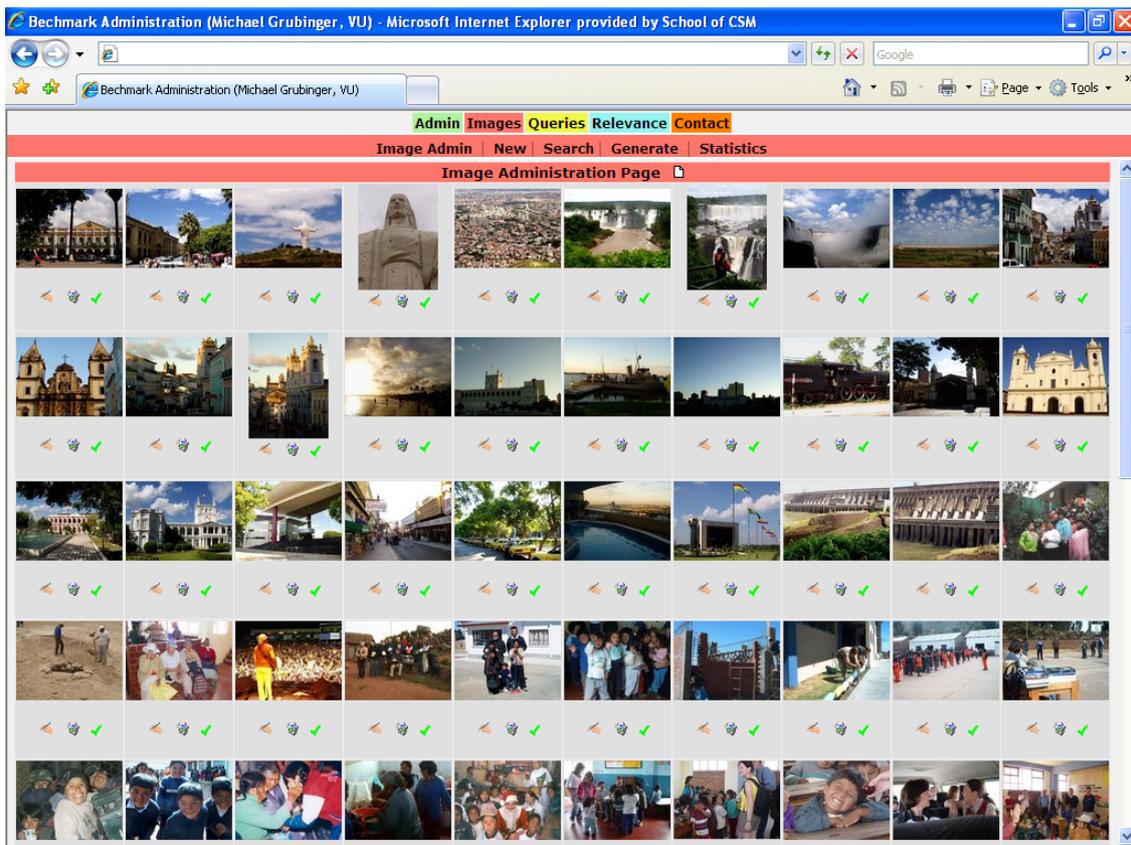


Figure 6.7: Image administration.

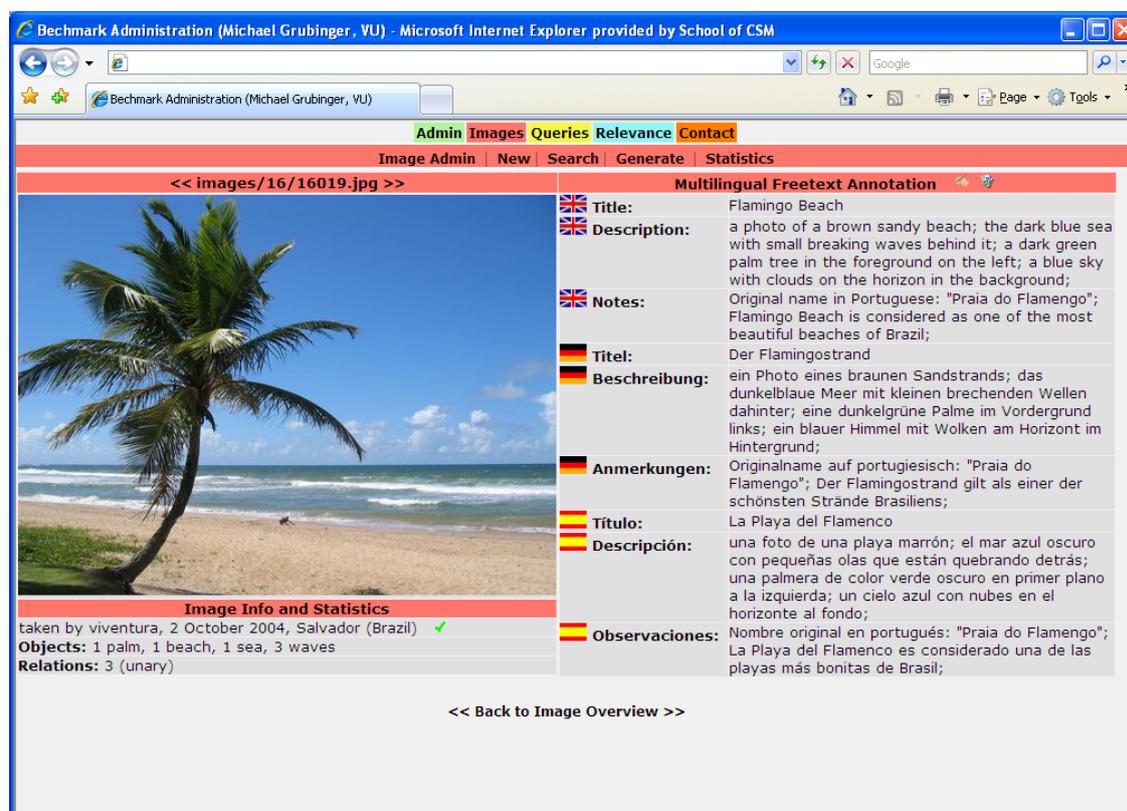
more significance with an increasing collection size, because the manual administration of the images is not feasible and effective anymore once the collection reaches a substantial number of image and representation files.

When an image is deleted, both image and thumbnail files, the entire corresponding information (meta-data and logical image representations) stored in the database, as well as all associated text files, are removed. As for the status indicator, a green tick means that this image has been annotated completely and that it can be distributed in a release status, whereas a red cross would mean that it should not yet be distributed.

If the mouse is moved over a thumbnail, the system displays the filename of that image, and if a thumbnail is clicked on, the system provides all the information of that particular image together with its semantic representation (this page is also the starting point for the management of the textual representation of an image).

Representation Management

This feature provides the necessary functions for the efficient, systematic and consistent creation of the logical image representations. Currently, images can be represented in a semi-structured format and in several languages including English, German and Spanish, and the functionality for the creation of keyword representations describing the major image objects and relationships is also provided.



The screenshot displays a web browser window titled "Benchmark Administration (Michael Grubinger, VU) - Microsoft Internet Explorer provided by School of CSM". The browser's address bar shows "Google". The page has a navigation menu with tabs: "Admin", "Images", "Queries", "Relevance", and "Contact". Below this is a sub-menu: "Image Admin", "New", "Search", "Generate", and "Statistics". The main content area is titled "<< images/16/16019.jpg >>". On the left, there is a large image of a palm tree on a beach. Below the image, the "Image Info and Statistics" section shows: "taken by viventura, 2 October 2004, Salvador (Brazil) ✓", "Objects: 1 palm, 1 beach, 1 sea, 3 waves", and "Relations: 3 (unary)". On the right, the "Multilingual Freetext Annotation" section provides descriptions in multiple languages:

Language	Title	Description	Notes
English	Title: Flamingo Beach	Description: a photo of a brown sandy beach; the dark blue sea with small breaking waves behind it; a dark green palm tree in the foreground on the left; a blue sky with clouds on the horizon in the background;	Notes: Original name in Portuguese: "Praia do Flamengo"; Flamingo Beach is considered as one of the most beautiful beaches of Brazil;
German	Titel: Der Flamingostrand	Beschreibung: ein Photo eines braunen Sandstrands; das dunkelblaue Meer mit kleinen brechenden Wellen dahinter; eine dunkelgrüne Palme im Vordergrund links; ein blauer Himmel mit Wolken am Horizont im Hintergrund;	Anmerkungen: Originalname auf portugiesisch: "Praia do Flamengo"; Der Flamingostrand gilt als einer der schönsten Strände Brasiliens;
Spanish	Título: La Playa del Flamenco	Descripción: una foto de una playa marrón; el mar azul oscuro con pequeñas olas que están quebrando detrás; una palmera de color verde oscuro en primer plano a la izquierda; un cielo azul con nubes en el horizonte al fondo;	Observaciones: Nombre original en portugués: "Praia do Flamengo"; La Playa del Flamenco es considerado una de las playas más bonitas de Brasil;

At the bottom of the page, there is a link: "<< Back to Image Overview >>".

Figure 6.8: Logical image representation management.

Figure 6.8 displays the main page of the submodule for the administration of the logical image representations. This page is composed of two separate parts.

The left part displays the particular image, its full path within the collection, and navigation arrows to go back to the last image or forward to the next one respectively. It further provides relevant information associated with that image, including its originator, its date of creation, the location and country where it was taken, its status in the collection (again, either a green tick or a red cross), and

further information on the objects and relationships within that image.

The right part displays the title, description and notes fields of the logical image representations in English, German and Spanish. If the edit symbol in the table header is clicked, the system provides a form to edit both the image data and the corresponding semantic descriptions (see Figure 6.9).

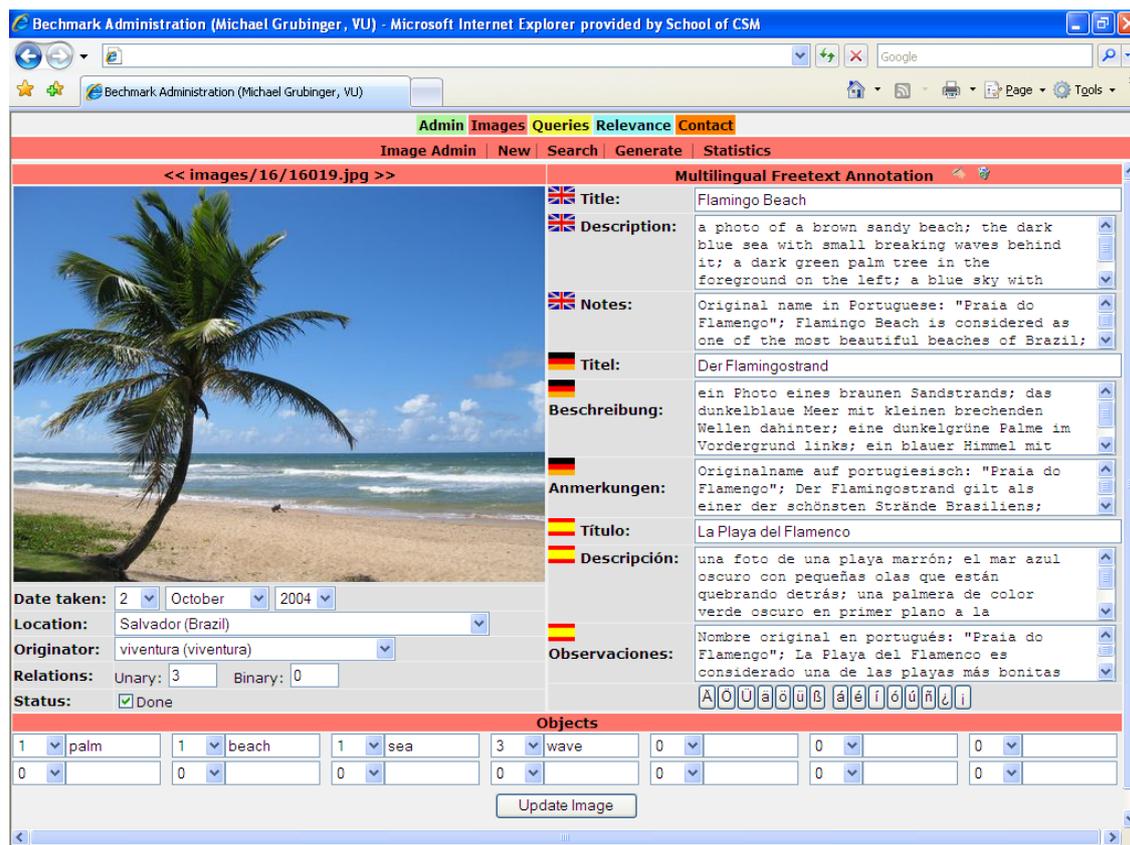


Figure 6.9: Edit logical image representations.

This page is composed of three parts: the one on the left provides the functionality to edit the image data, the one on the right the functions to edit the semantic descriptions of the image, while the one at the very bottom facilitates the specification of the objects and their cardinality values within that image.

The system also provides buttons to add special symbols (that might or might not be available on every keyboard) to further facilitate the annotation effort. By pressing the “Update Image” button, the information in the database is updated and the system returns to the main image representation page (Figure 6.8).

Subset Generation

The parametric nature of the image benchmark makes it possible to create different subsets of the image collection. While such parameterisation of a test collection provides an abundance of benefits for the organisers of an evaluation event (as indicated above), it would be rather counterproductive to directly offer the same functionality to its participating groups as well, because this could potentially create a similar situation as experienced with the *Corel* collections: researchers using different subsets to highlight their own algorithm’s benefits (see Section 3.2.1, and in particular [294]).

In order to avoid such a dilemma (and possible cheating) in an evaluation event, it seems more reasonable to provide the participants with identical and static subsets of the document collection (and to keep the parameters used for the creation of those, which is vital for the reproducibility of that particular evaluation event). Hence, we also implemented such an export mechanism which allows the generation of image subsets with respect to the specified parameters.

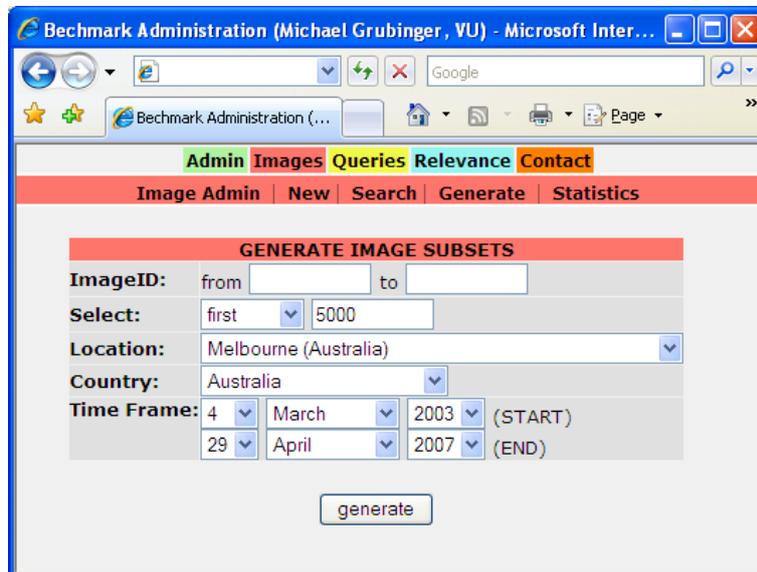


Figure 6.10: Collection subset generator.

Figure 6.10 illustrates the page for the generation of image subsets, which can be accessed by choosing the “Generate” option in the “Images” submenu. The cur-

rent implementation of the system thereby allows the specification of the following parameters to generate subsets with respect to:

- the unique image identifiers (*e.g.* images with an ID between 1000 and 2000);
- the images' position in the database (*e.g.* the first/last/random 5000 images);
- the image locations (*e.g.* only images from Melbourne);
- the country of image creation (*e.g.* only images taken in Australia);
- the time frame of image creation (*e.g.* images taken between 2002 and 2004);
- any combination of these parameters.

The resulting subsets can either (1) be downloaded via a link provided by the system or (2) be stored in a predefined directory on a server accessible by the system.

Like the creation of image subsets, the generation of the associated logical image representations can be varied with respect to several parameters as well.

Figure 6.11 displays the module used to export the semantic representations of the images within the collection. In particular, it shows the settings used for the majority of files exported for *ImageCLEFphoto* 2006: complete multilingual image representations in a semi-structured format as specified by CLEF, in English and German, with all fields provided, and 100% orthography, *i.e.* no spelling mistakes injected (see also Section 7.2.2).

This submodule can thereby be accessed in the same way as the image generation submodule. Image representation parameters include:

- the representation type (*e.g.* multilingual free-text representations);
- the representation format (*e.g.* as used in CLEF);
- the representation language (either English, Spanish, German, or with a randomly selected language for each image);

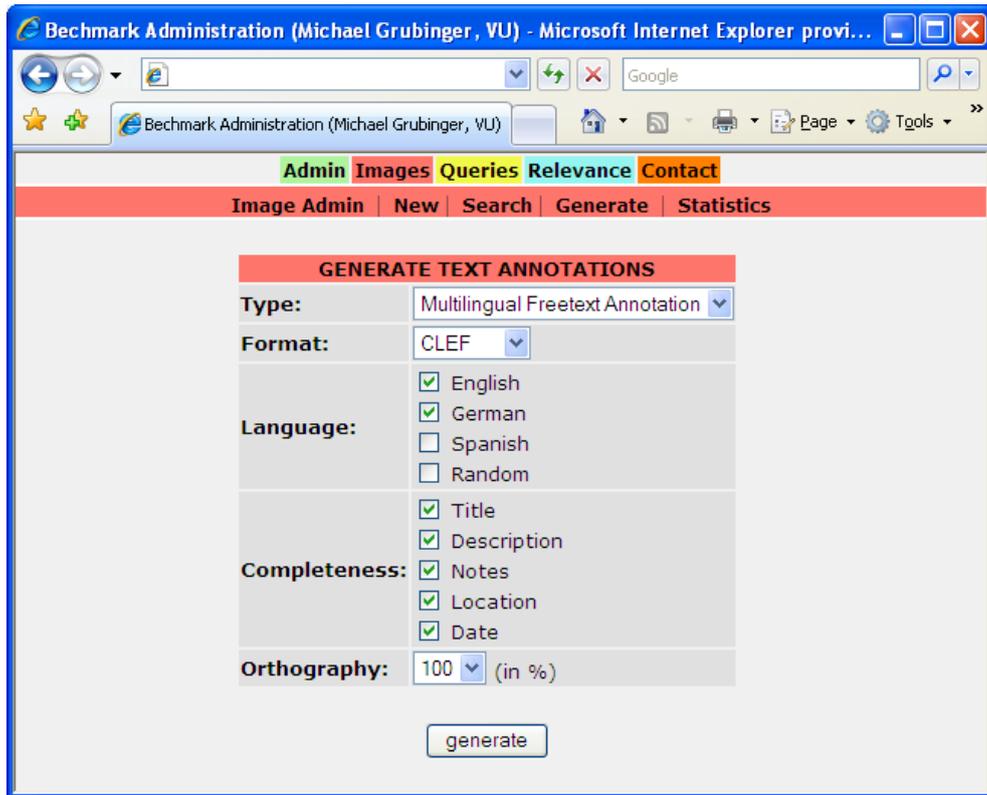


Figure 6.11: Image representation generator.

- the representation completeness (title, description, notes, location and date fields can either be selected or unselected);
- the level of orthography (100% means no errors are introduced, 0% that a spelling mistake is injected in every single word);
- any combination of these parameters.

The result of the generation is simple text files, which can either (1) be directly downloaded via a link offered by the system, or (2) be stored in a predefined directory on a server accessible by the system.

Miscellaneous

The submenu of the collection management module contains two further options, *Search* and *Statistics*. Both have not yet been implemented, but it is planned to start their implementation in the near future.

6.3.3 Topic Management and Administration

The topic creation process for the *IAPR TC-12 image collection* was certainly not a trivial task and involved the consideration of several dimensions (see Section 5.4.2). As a consequence, we developed a module to guide the creation process for these query topics and to provide the necessary functionality to facilitate their subsequent management and analysis as well as their versatile generation to text files.

This section introduces the module for topic management and administration, which constitutes another key component of our benchmark administration system. If this module is selected, the system displays the topic overview page by default.

Topic Overview

Figure 6.12 illustrates the topic overview page, which shows the topic titles ordered by their unique identifiers and eight columns next to each topic.

Admin Images Queries Relevance Contact									
Overview New Search Generate Print Statistics									
Query Administration Page									
ID	Title	TS	QD	QL	VL	LF	Sh.	Ed.	Del.
1	accommodation with swimming pool	37	2	4	3	✓	✱	✱	✱
2	church with more than two towers	17	4	6	3	✓	✱	✱	✱
3	religious statue in the foreground	37	3	5	1	✗	✱	✱	✱
4	group standing in front of mountain landscape in Patagonia	28	4	9	2	✓	✱	✱	✱
5	animal swimming	45	2	2	3	✓	✱	✱	✱
6	straight road in the USA	40	3	5	3	✓	✱	✱	✱
7	group standing in salt pan	40	3	5	5	✓	✱	✱	✱
8	host family posing for a photo	70	2	4	1	✓	✱	✱	✱
9	tourist accommodation near Lake Titicaca	14	4	5	1	✓	✱	✱	✱
10	destinations in Venezuela	45	2	3	1	✓	✱	✱	✱
11	black and white photos of Russia	56	2	6	5	✗	✱	✱	✱
12	people observing football match	35	3	4	3	✓	✱	✱	✱
13	exterior view of school building	67	2	5	2	✓	✱	✱	✱
14	scenes of footballers in action	41	2	6	4	✓	✱	✱	✱
15	night shots of cathedrals	29	2	4	5	✗	✱	✱	✱
16	people in San Francisco	36	2	4	1	✓	✱	✱	✱
17	lighthouse at the sea	28	2	4	3	✓	✱	✱	✱
18	sport stadium outside Australia	49	4	4	3	✗	✱	✱	✱
19	exterior view of sport stadium	36	2	4	3	✗	✱	✱	✱
20	close-up photograph of an animal	93	2	5	3	✓	✱	✱	✱
21	accommodation provided by host families	70	3	5	2	✓	✱	✱	✱
22	tennis player during rally	98	2	4	4	✗	✱	✱	✱
23	sport photos from California	82	2	4	1	✓	✱	✱	✱
24	snowcapped building in Europe	21	3	4	3	✓	✱	✱	✱
25	people with a flag	35	3	4	1	✓	✱	✱	✱
26	godson with baseball cap	46	2	4	3	✓	✱	✱	✱

Figure 6.12: Topic overview.

The first five columns thereby contain further information on each of the topics: the estimated size of the target set (TS), the linguistic difficulty of the query (QD), the query topic length (QL), its level of visuality (VL) and whether a topic had been taken (or derived) from the log file or not (LF). The topics can also be sorted according to these dimensions by clicking on the column headings, while moving the mouse over these headings would spell out their abbreviations.

The last three columns contain the already introduced and clickable symbols to edit, delete or show the information of a topic, while the table heading contains the symbol to add a new topic.

Topic Creation

One of the most significant subcomponents of the topic management module is the feature that facilitates and guides the topic creation process to add new topics to the database (see Figure 6.13).

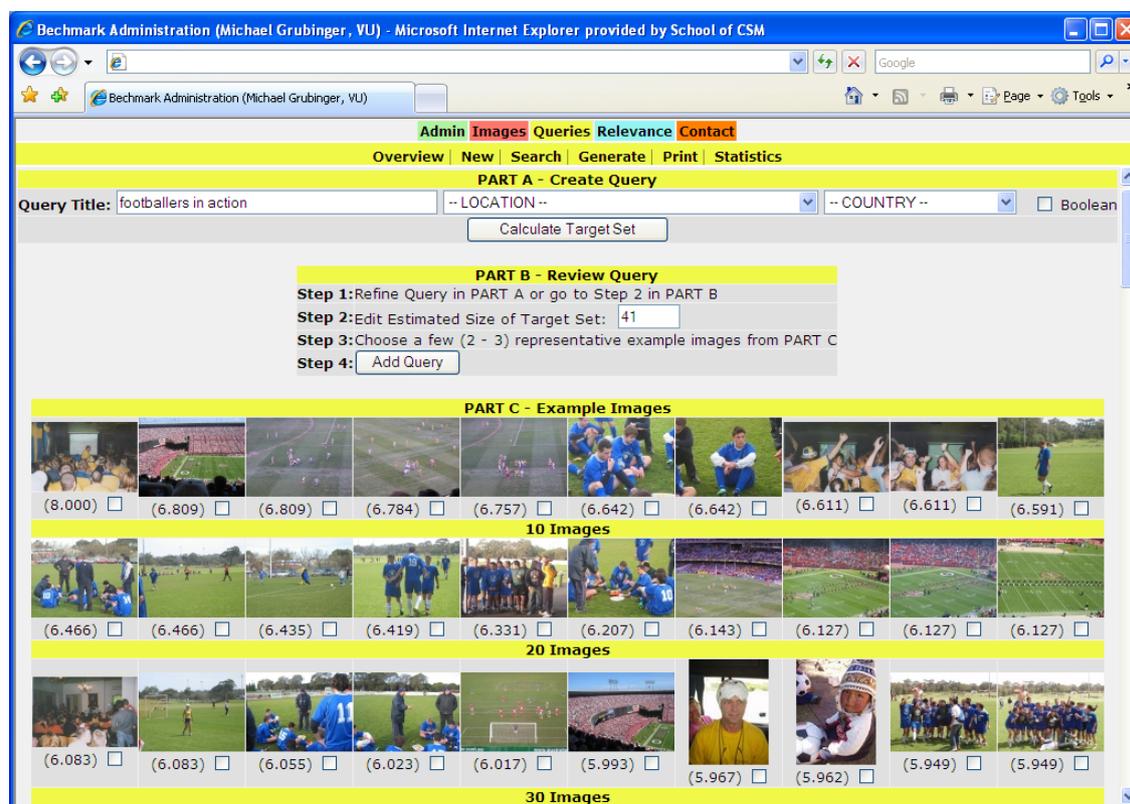


Figure 6.13: Topic creation.

This page can be accessed by selecting the “New” option in the yellow sub-menu, or by clicking the document symbol in the topic overview. The creation of a representative query topic is thereby broken down into several steps:

1. In the first step (Part A), the system allows the user to *create a query* and subsequently searches the document collection, returns a list of potentially relevant images and displays their thumbnails. To assist with finding representative topics, the location and country of relevant images can be specified, and boolean searches can be used as well.
2. The second step (Part B) offers a function to *review the query*: the user can either (1) *refine* it to start the creation process again, or (2) continue and redefine the number of relevant images for that particular query. The system automatically shows a number of candidate images that are ordered by decreasing relevance according to a simple text-matching algorithm using BM25.
3. In the third step (Part C), sample images can be selected which will provide the basis for the QBE paradigm used for CBIR approaches, and the images relevant to the query can be identified and marked to establish a preliminary ground-truth.
4. Finally, the new query topic is added to the database and the system shows the information page of the newly created topic.

Topic Administration

Once a query topic is added to the database, it can be further completed, edited or deleted - which are the main functionalities provided by the topic administration feature. Figure 6.14 presents the topic information page, which can be accessed by clicking on the symbol of the yellow star in the topic overview page.

The first line in yellow states the event in which the topic is used, displays the unique identifier of that topic within that event and provides the symbols to

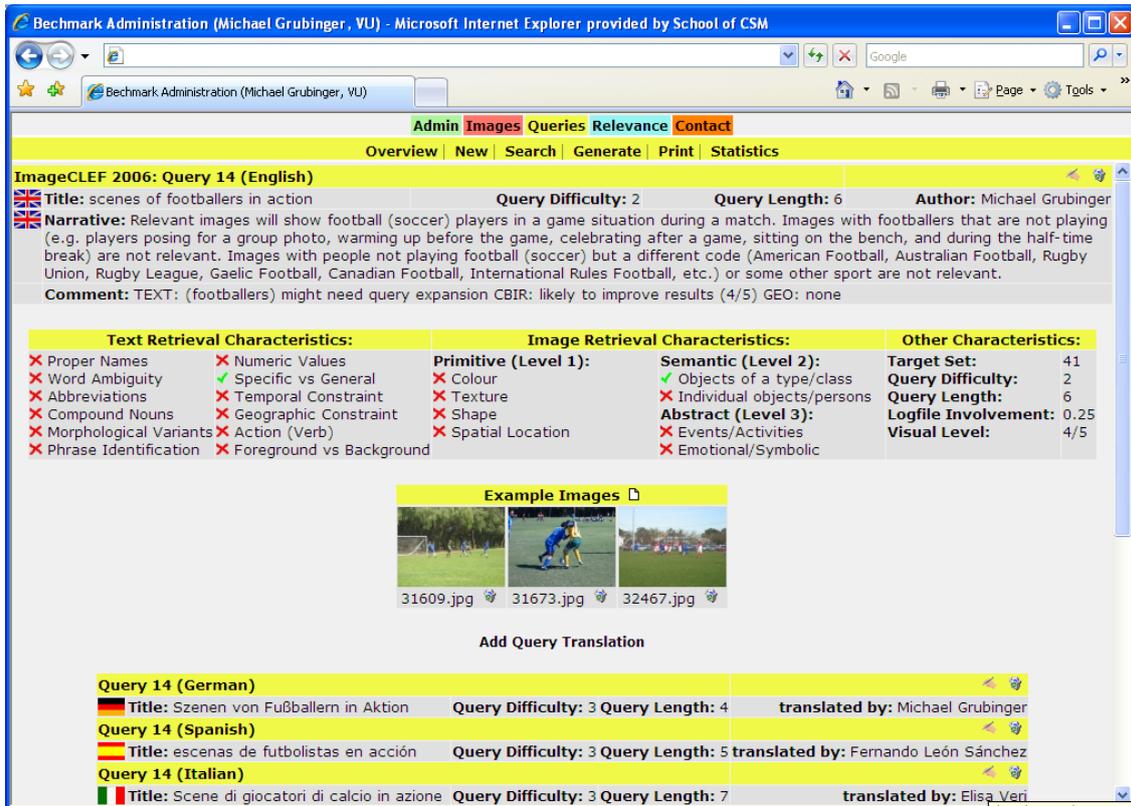


Figure 6.14: Show topic information.

further edit or delete this topic. General information on this topic, such as the title, narrative, difficulty, length and author of the topic as well as comments on the topic, is shown. Both text and image retrieval characteristics are indicated, and further statistical data are given. Sample images for this topic are shown and the functionality to further add or delete these images is provided, together with an overview of all existing translations for that particular topic and the link to add further translations.

In order to edit a topic, the system provides a form (see Figure 6.15) which allows the user to carry out changes for all these aforementioned fields. By clicking the “Edit Query” button, the changes are committed in the database and the system returns to the information page of that particular topic. If a topic is deleted, the system returns to the topic overview page.

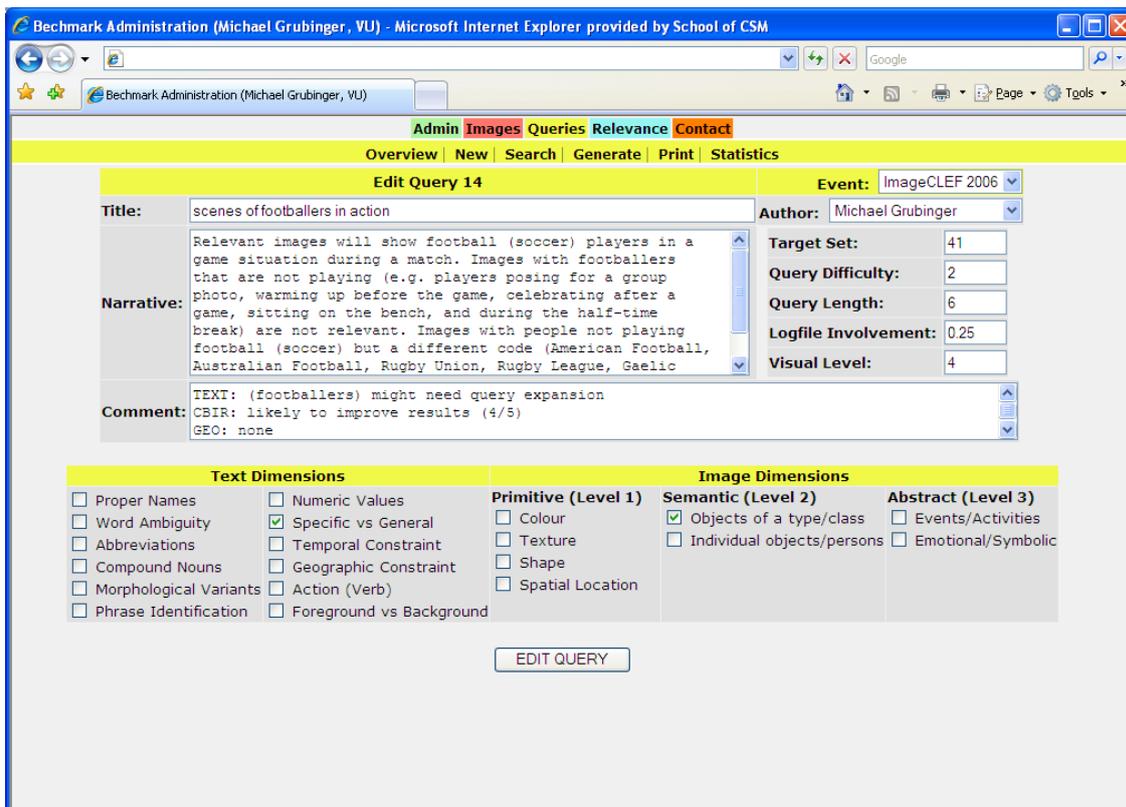


Figure 6.15: Edit topic.

Topic Translation

The topic overview page also provides access to another essential feature within the topic management application: the translation of the query titles and narratives, which is a vital component for any multilingual evaluation environment. The system currently supports 15 topic languages which are predominantly used at *ImageCLEF* (including English, German, Spanish, Italian, French, Portuguese, Dutch, Norwegian, Swedish, Finnish, Danish, Polish, Russian, Chinese and Japanese), but can easily be expanded to accommodate more languages.

Figure 6.16 illustrates the bottom part of the topic information page, which does not only display topic specific information but also serves as the starting point for the translation of that particular topic. The topic translation overview indicates the title, linguistic difficulty, length and author for each translation of the topic and further provides the functionality to add, edit and delete these translations.

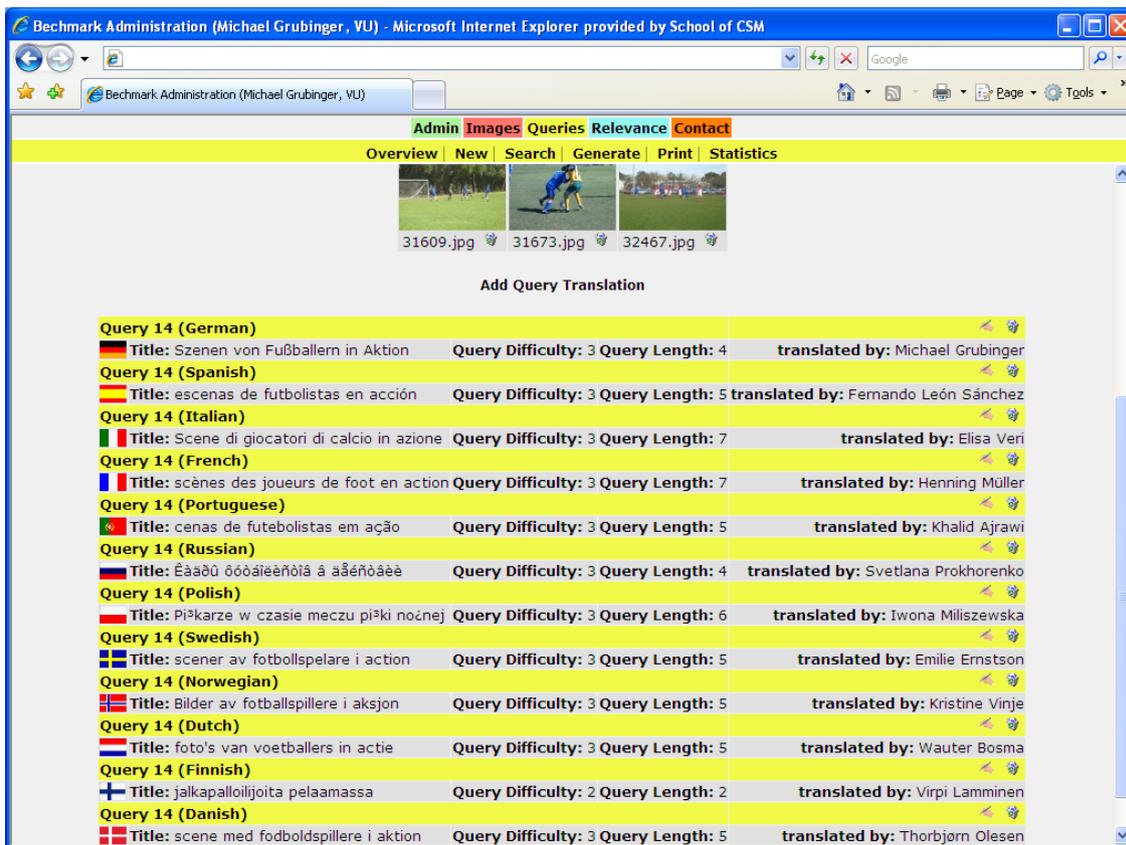


Figure 6.16: Topic translation overview.

A topic translation can be added by clicking on the “Add Topic Translation” link, which is located above the topic overview. The system then provides an empty form (see Figure 6.17) to perform the translation into any language that has not yet been used for this particular topic.

In order to edit a topic, the system responds with a translation edit form as shown in Figure 6.18.

This form allows the changes of the title, narrative, author and linguistic difficulty values of the topics; only the topic language (indicated by the heading and the flag) is fixed and cannot be changed. If a topic translation is deleted, it is permanently removed from the database and also from the overview.

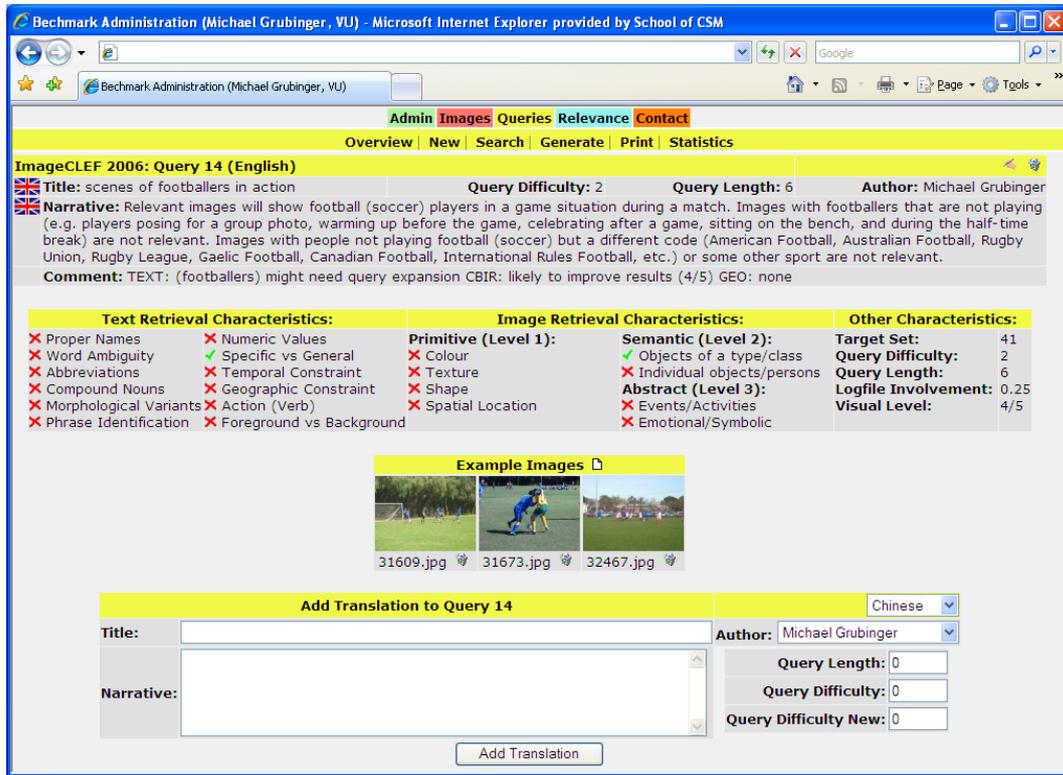


Figure 6.17: Add topic translation.

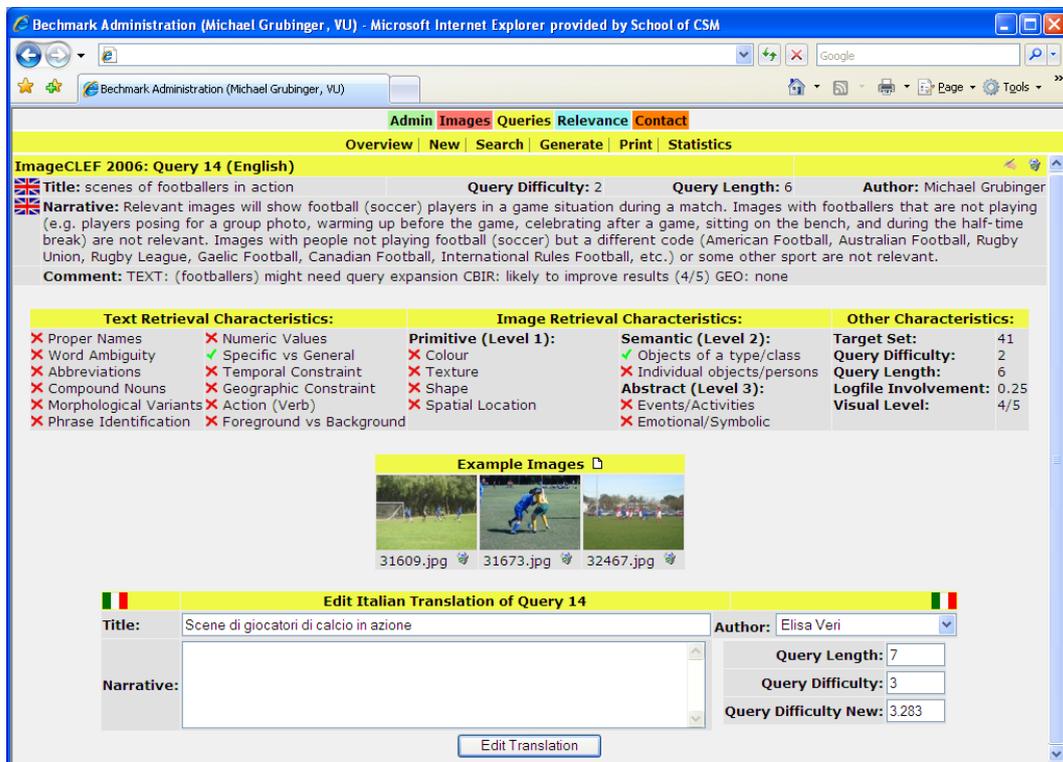


Figure 6.18: Edit topic translation.

Topic Generation

Like the collection management module, the topic management module also provides an export function that facilitates the automatic generation of the topics and allows for their subsequent distribution to the participants.

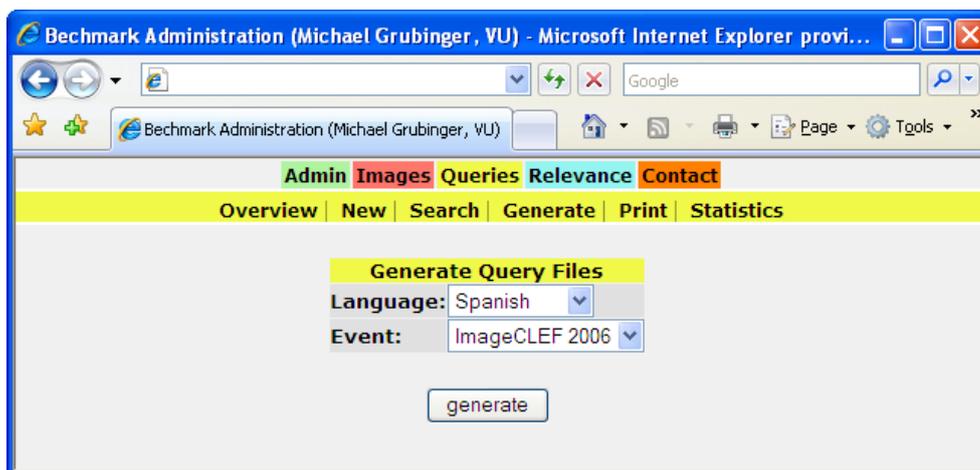


Figure 6.19: Topic generation.

As indicated in Figure 6.19, this module is not as parameter intensive as the one in the collection management system, with the only parameters being the topic format (indicated by the event that the topics are generated for) and the respective topic language.

The resulting topics can either (1) directly be downloaded via a link provided by the system or (2) be accessed in a predefined directory on a server accessible by the system.

Print Topics

The topic administration system also provides the functionality to print all the topics of one language and event on one page (see Figure 6.20).

This page, which can be accessed by selecting the “Print” option in the “Queries” submenu, displays the topic titles, narratives and all sample images (ordered by Topic ID).

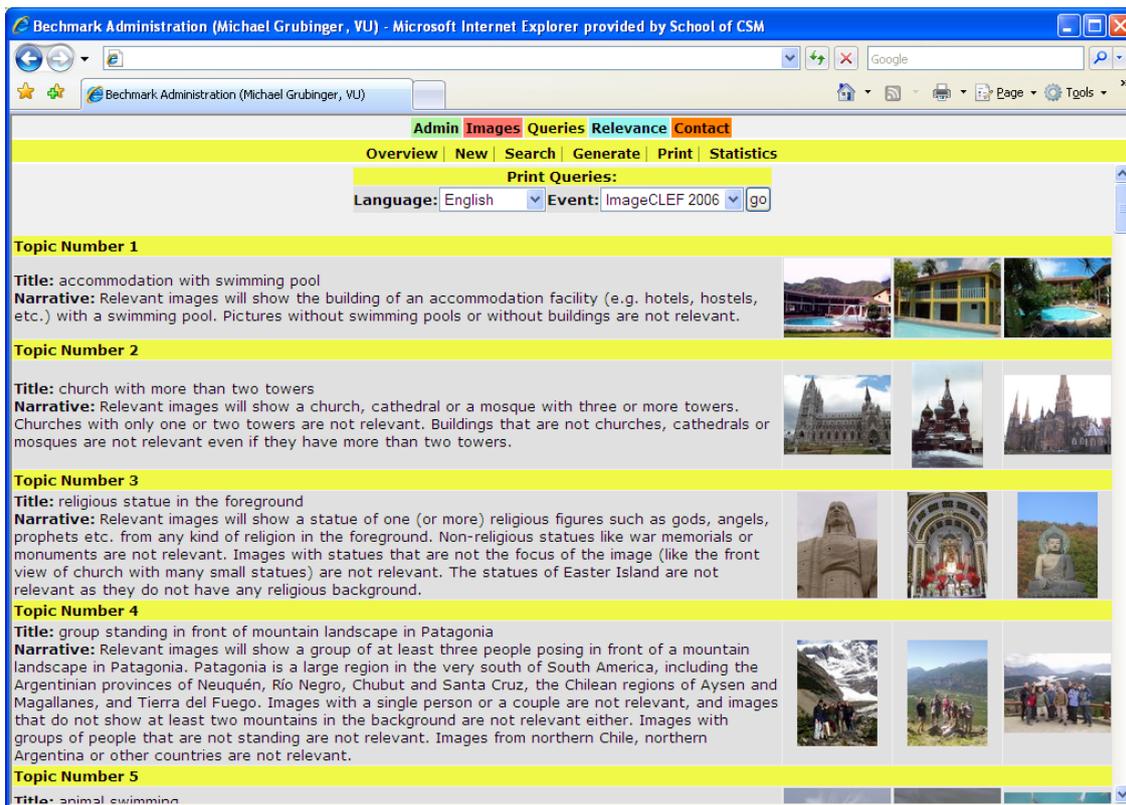


Figure 6.20: Print topics.

6.3.4 Relevance Assessment

The blue link on the main page, *Relevance*, leads to the relevance assessment module of the benchmark administration system⁶.

The implementation of this module is based on a text indexing system that ranks images using the *BM25* weighting operator. No stemming or query processing is performed, and only a basic list of stop-words (similar to the one provided with the SMART system [33]) is used.

The module itself is written in Perl and uses the *common gateway interface* (CGI) and Perl templates for the user interface, while the text indexing system for ISJ is implemented in C.

⁶This module was provided courtesy of Paul Clough and Mark Sanderson, University of Sheffield, UK.

Relevance Assessments Overview

This module first offers the selection of an event that is supported by the system and then displays an overview page once a particular event has been chosen (see Figure 6.21).

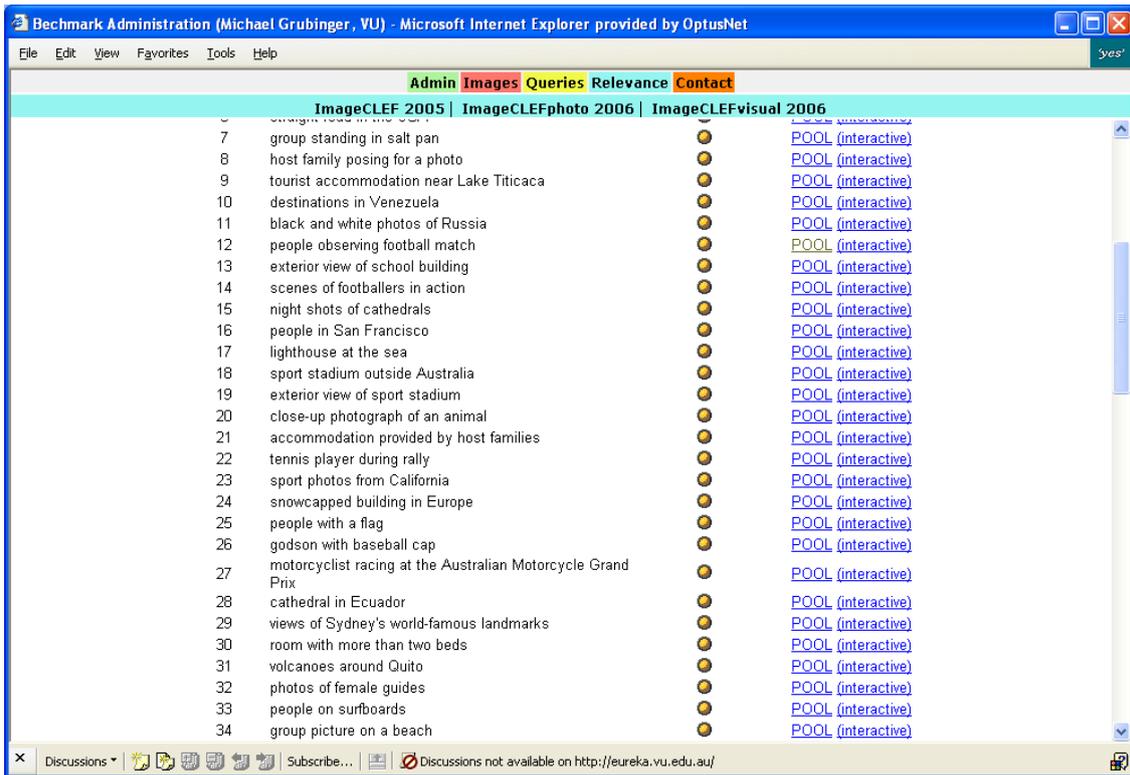


Figure 6.21: Relevance assessment overview.

This page shows all the topics of the chosen event ordered by their unique topic identifier and provides links to carry out pooled relevance assessments (POOL) and interactive relevance assessments (interactive) for each of the topics. The yellow sign next to the topic indicates whether this particular topic has already been judged or not.

Pooled Relevance Assessments

Figure 6.22 displays the page to carry out the pooled relevance assessments, showing an example of the judgment for the sample topic “scenes of footballers in action”.



Figure 6.22: Pooled relevance assessments.

The system displays the topic title, the narrative descriptions and two sample images on the top of the screen (with a blue background). Below it, it shows all the images from the candidate pool for that particular topic (877 in the example above), which are ranked by decreasing order of the percentage of systems that agreed on that image being relevant to make use of the benefits of MTF pooling (see Section 3.4.2), whereby its limitations are outweighed by the fact that all the images in the pool have been judged by the topic assessors (see Section 7.2.5). The thumbnail (which can be clicked to display the large version of the image) as well as the logical image representations are shown, and the ternary judgments scheme is offered.

The relevance assessments are carried out for each image in the pool by selecting one of the three options: (1) relevant, (2) partially relevant and (3) not relevant. The default setting thereby is “unjudged”; any image that has not been judged is considered as irrelevant by default. Furthermore, it is also possible to undo the

relevance judgment for an image by clicking the “Remove judgment” button.

Pressing the “Return to Topics” button finally lets the user return to the relevance assessments overview page.

Interactive Relevance Assessments

Figure 6.23 illustrates the page for the interactive relevance assessments which are used to complement the judgments based on the pooled relevance assessments.



Figure 6.23: Interactive relevance assessments.

The user interface is very similar to that used for the pooled relevance assessment, with one exception: the system also offers the assessor a concept-based search interface which facilitates the use of ISJ to find as many further relevant images as possible for the particular topic in question. This ISJ functionality was also used to determine the alternative low-cost methods for the estimation of topic difficulty presented in Section 5.3.2.

6.3.5 Contact

Finally, the contact section provides information on the main researchers working on this project: Professor Dr. Clement H. C. Leung (Figure 6.24) and Michael Grubinger (Figure 6.25).



Figure 6.24: Contact information Clement Leung.



Figure 6.25: Contact information Michael Grubinger.

6.4 Summary

This chapter introduced a novel parametric benchmark architecture and administration system. This comprises the identification of various essential benchmark

parameters with respect to the image collection, the corresponding semantic descriptions of the images and the representative query topics. Based on these parameters, we derived a physical database model and implemented a parametric benchmark administration system to facilitate and guide the management of the major benchmark components as well as to enable a deeper understanding of the complex processes associated with the preparation and organisation of an evaluation event for VIR.

The most significant benefit of the architecture presented in this chapter is its parametric nature, which allows for the fast adaptation to changed retrieval requirements or new evaluation needs. This parametric benchmark paradigm is supported by our benchmark administration system, which we specifically designed and implemented to allow the quick reaction to such changes in direction by simply adjusting the parameters and the subsequent regeneration of the required subsets. This is a major advantage over the static collections used in many other evaluation events, whereby new data collections often have to be acquired to react to changes, which can be a very time consuming and cost-intensive task. Further merits of the benchmark administration system include:

- the facilitation of the incremental collection development;
- the guidance of the topic creation, management and analysis processes;
- the administration of the topic translations;
- the efficient execution of relevance assessments.

The creation of the *IAPR TC-12 Image Benchmark*, including the freely available image collection together with a set of representative query topics and a predefined ground-truth, all of which would have not been possible without the use of the parametric benchmark administration system, has certainly made a significant contribution to the field of VIR. However, a benchmark can only be beneficial if researchers can also be motivated to make use of them in evaluation events. In the next chapter, we therefore report on the first evaluation event for ad-hoc retrieval from a generic photographic collection (*ImageCLEFphoto 2006*).

Chapter 7

System Analysis and Evaluation

The previous chapters described the design and development of a parametric administration architecture for the *IAPR TC-12 Image Benchmark*: a test collection for VIR comprising an image collection of generic photographs, a set of representative query topics and a predefined ground-truth associated with each of them. Although this benchmark provides excellent resources to the information retrieval and computational vision communities to facilitate the standardised laboratory-style testing of (predominately concept-based) image retrieval systems, it would only prove beneficial to research if its components were actually used in evaluation events as well. Therefore in this chapter, we report on the involvement of the *IAPR TC-12 Image Benchmark* at *ImageCLEFphoto 2006*: the first evaluation event for (multilingual) ad-hoc retrieval from generic photographic collections.

Section 7.1 first presents an introduction to ad-hoc retrieval tasks at *ImageCLEF* and states the reasons for the choice of a multilingual environment as evaluation platform. Section 7.2 then introduces the validation design and describes the organisation and realisation of *ImageCLEFphoto 2006*. Section 7.3 concentrates on the description of the retrieval systems and provides an analysis of their performance according to several submission parameters and topic dimensions. Section 7.4 first quantifies the quality of the benchmark and then provides an analysis of the event itself, which includes the evaluation of the task difficulty, the choice of performance measures, participants' feedback and based on it, the future prospects of this task. Parts of this chapter are taken from [60, 61].

7.1 Introduction

This section provides an introduction to ad-hoc retrieval evaluation at *ImageCLEF* and presents the motivation for the involvement of the *IAPR TC-12 Benchmark* within the general ad-hoc retrieval task offered by *ImageCLEF*.

7.1.1 Ad-hoc Retrieval Evaluation at ImageCLEF

ImageCLEF was established in 2003 with the aim of evaluating content and concept-based image retrieval from multilingual document collections and has since offered a variety of tasks for both system-centred and user-centred retrieval evaluation within two main areas: retrieval of images from photographic collections and retrieval of images from medical collections. These fields have helped to attract different groups to *ImageCLEF* (and CLEF) and to broaden the audience of this evaluation campaign (see also Section 3.6.2).

One of the key tasks of *ImageCLEF* is concerned with evaluation of system performance for ad-hoc image retrieval from photographic collections in a laboratory-style setting. This kind of evaluation is system-centred and 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 the search topics are not known to the system in advance. Evaluation thereby only concentrates on comparing algorithms and systems and does not aim to assess aspects of user interaction as such evaluation is carried out in other tasks such as [136, 137]. The specific goal of the ad-hoc retrieval task is: given an alphanumeric statement (and/or sample images) describing a user information need, find as many relevant images as possible from the given collection (with the query language either being identical or different from that used to describe the images).

From 2003 to 2005, the general ad-hoc retrieval task was based on cross-language retrieval from a cultural heritage collection: the SAC of historic photographs (see Section 3.2.2). This provided certain challenges for both the text and visual retrieval communities, most noticeably the style of language used in the logical image

representations and the types of pictures in the collection: mainly black-and-white of varying levels of quality and visual degradation [62, 63, 64].

In 2006, the SAC was replaced by the *IAPR TC-12 Image Benchmark*, and the general ad-hoc retrieval task from photographic collections was given a new name (*ImageCLEFphoto*) in order to avoid confusion with the medical ad-hoc retrieval task (*ImageCLEFmed*).

7.1.2 Motivation

The involvement of the *IAPR TC-12 Benchmark* as the main test collection for *ImageCLEFphoto* has brought benefits for both sides. The main reasons why we approached *ImageCLEF* and offered the unconditional and free use of the *IAPR TC-12 Benchmark* as well as our manpower for organising *ImageCLEFphoto 2006* include the following:

- *ImageCLEF* had been a well established evaluation event since 2003 and its ad-hoc retrieval task was already offering a very similar task scenario to that modelled by the *IAPR TC-12 Benchmark* as well. Hence, we felt that it would be more sensible to combine forces and approach *ImageCLEF*, rather than creating yet another evaluation event offering a similar task in competition with *ImageCLEF* and thus further splitting this field of research.
- *ImageCLEF* had been attracting a large number of different research groups from several fields of IR including cross-language information retrieval (CLIR), CBIR and TBIR, and we therefore expected (and hoped for) a satisfactory level of participation for retrieval evaluation from the *IAPR TC-12 Benchmark* as well.
- *ImageCLEF* had provided a multilingual evaluation environment, which in the case of evaluation of retrieval from generic photographic collections represents the most realistic model as such real-life collections (especially online photo collections such as *FlickrR*) are inherently multilingual.

- We did not have the resources to organise an evaluation event on our own in order to apply the *IAPR TC-12 Benchmark* in practice.

For *ImageCLEF*, on the other hand, after three years of evaluation using the SAC of historic photographs, the move to a novel test collection for the standard ad-hoc retrieval task was motivated by several reasons, including the following:

- Based on feedback from *ImageCLEF* participants in 2004 and 2005, the organisers had noticed some saturation of interest in using SAC again for evaluation: mainly black and white images as well as varying levels of quality and visual degradation limited the successful use of CBIR methods, and concerns had been raised whether research results achieved within the limited domain of historic images would be transferable to more general retrieval situations.
- The *IAPR TC-12 Benchmark* was specifically designed as a benchmark collection, and it was considered as very well-suited for the use in *ImageCLEF*, with logical image representations in multiple languages and high-quality colour photographs covering a wide range of topics.
- *ImageCLEF* had always been focussing on realistic applications, and this type of collection - generic photographs - was estimated to be likely to become of increasing interest to researchers with the growth of the desktop search market and the popularity of tools such as *Flickr*.
- A similar logical image representation format as used with the SAC would offer a smooth transition for participants from the previously used SAC to the new test collection (*e.g.* to keep changes in existing retrieval and evaluation scripts to a minimum).
- One of the biggest factors influencing which collections are used and provided by *ImageCLEF* is copyright: the *IAPR TC-12 Benchmark* is available royalty-free, and no copyright restrictions hinder the large-scale redistribution of the collection to registered participants.

After the *IAPR TC-12 Image Benchmark* had been presented at the *ImageCLEF* workshop in 2005, both participants and organisers unanimously decided for its use in the general ad-hoc retrieval tasks from 2006 on.

7.2 Evaluation Design and Organisation

This section introduces the evaluation design of *ImageCLEFphoto 2006* and reports on the organisational aspects and the actual realisation of this evaluation event. Based on a slightly adapted model of the TREC-style benchmark (compare Figure 3.1 in Section 3.1.3), we decided to design the following chronological evaluation architecture.

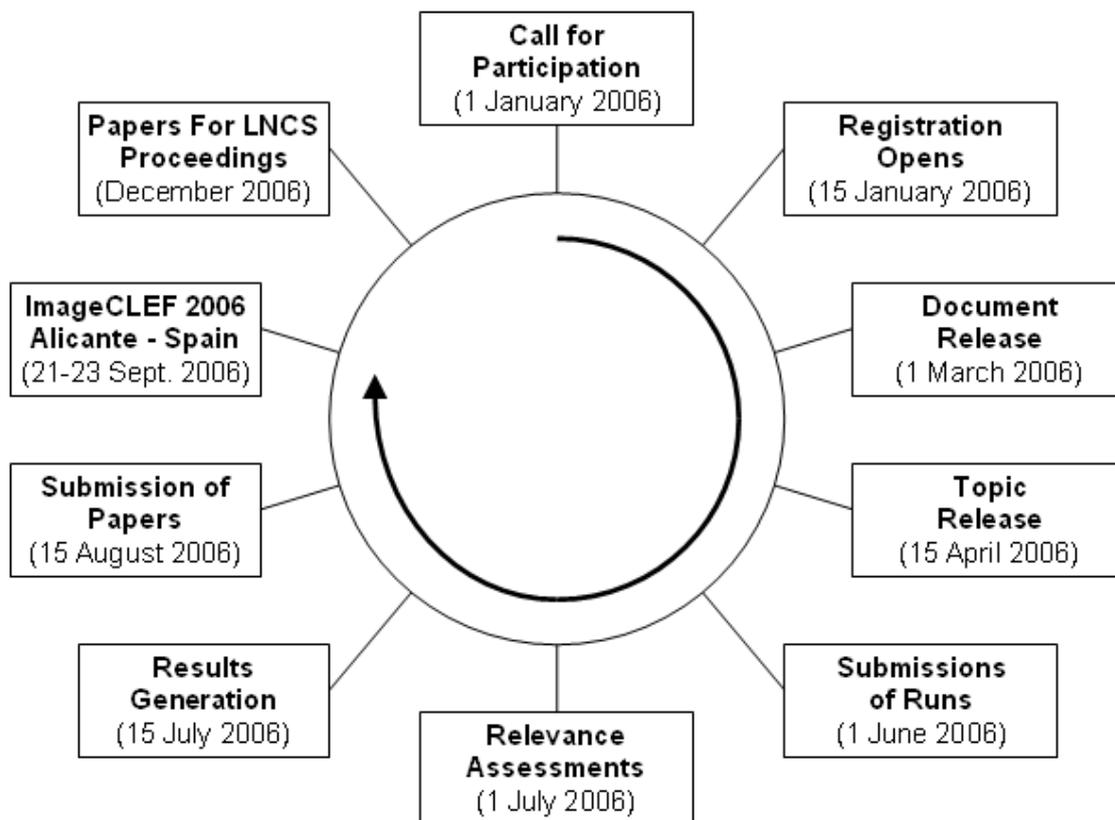


Figure 7.1: The annual cycle of *ImageCLEFphoto 2006*.

Figure 7.1 illustrates the cycle of events followed by *ImageCLEFphoto 2006*, together with the corresponding time frame of the event. Each individual component will be further discussed in chronological order below.

7.2.1 Call for Participation and Registration

ImageCLEFphoto 2006 officially started in early January 2006 with a *call for participation*, in which we first presented the novel tasks and challenges for the planned evaluation event and encouraged research groups to participate. This included the production of flyers and the creation of an *ImageCLEF 2006 web site*¹ to provide an information platform about the event and to further promote it by means of both online and offline media. We also set up an *ImageCLEF* mailing list which we used to inform past *ImageCLEF* participants as well as new research groups that had showed interest in participating.

The call for participation encouraged researchers to use any method they wished for to retrieve relevant images, especially the use of combined concept-based and content-based retrieval methods. Further key research areas that were addressed include the investigation of:

- various methods of query translation;
- how features derived from the images and their logical representations could be combined to enhance retrieval;
- how text and image attributes could be combined to enhance cross-language image retrieval in this kind of domain;
- how vocabulary mismatches between the logical image representations and queries could be bridged.

Registration then opened on 15 January 2006, and all prospective participants had to sign a registration and a data release form to officially register for CLEF and to gain access to the test collections. *ImageCLEFphoto 2006* saw the registration of 36 research groups from 21 different countries and 4 continents, indicating not only the success of the call for participation, but also the immense need for the evaluation of VIR from generic photographic collections and the global interest of

¹<http://ir.shef.ac.uk/imageclef/2006/>

researchers world-wide to participate in such an evaluation event: 18 registrations were from Europe, eleven from Asia, six from America and one from Australia.

7.2.2 Document Release

After the SAC of historic photographs had been used for three years, the *IAPR TC-12 image collection* provided a novel database for evaluation in 2006. Unlike many existing photographic collections used to evaluate VIR systems, this collection is very generic in content, with many different images of similar visual content but varying illumination, viewing angle and background. This makes it a challenge for the successful application of techniques involving visual analysis (see Chapter 4).

Document Access and Distribution

Only registered participants were granted access to the entire document collection of 20,000 photographs (and 20,000 corresponding thumbnails) on 15 March 2006. In addition, using the image collection management system presented in Section 6.3, we exported the semantic descriptions from the database and generated text files in English and German. The resulting corpus of 80,000 files was organised into one single archive as described in Section 4.2.6 and was subsequently made available for download to the registered participants.

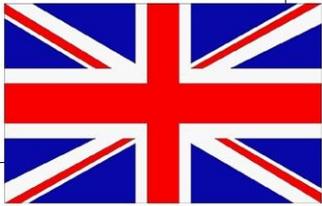
<pre><DOC> <DOCNO>annotations/16/16019.eng</DOCNO> <TITLE>Flamingo Beach</TITLE> <DESCRIPTION> a photo of a brown sandy beach; the dark blue sea with small breaking waves behind it; a dark green palm tree in the foreground on the left; a blue sky with clouds on the horizon in the background; </DESCRIPTION> <NOTES> Original name in Portuguese: "Praia do Flamengo"; Flamingo Beach is considered as one of the most beautiful beaches of Brazil; </NOTES> <LOCATION>Salvador, Brazil</LOCATION> <DATE>2 October 2002</DATE> <IMAGE>images/16/16019.jpg</IMAGE> <THUMBNAIL>thumbnails/16/16019.jpg</THUMBNAIL> </DOC></pre>	
--	--

Figure 7.2: The generated English caption for image 16019.jpg.

Figure 7.2 provides an example of the English representation for image 16019.jpg. We used the following parameters to generate the text files for *ImageCLEFphoto 2006*: representation type, format, language, completeness and the level of orthography (these settings correspond to the ones shown in Figure 6.11).

Representation Type and Format

As far as the type and format of the logical image representations are concerned, we decided to create semantic descriptions using a similar format to that of the SAC (see Section 3.2.2) used in previous years: multilingual text representations in a semi-structured format.

Thus, similar to the SAC, the entire representation was nested between the <DOC> and </DOC> tags, and the <DOCNO> tag contained the pathname of the text file as a unique document identifier, while the title, description, notes, location and date fields were represented by the <TITLE>, <DESCRIPTION>, <NOTES>, <LOCATION> and <DATE> tags. In addition, the <IMAGE> and <THUMBNAIL> tags contained the path of the actual image file and its corresponding thumbnail respectively (see Figure 7.2).

The choice for this format had been based on the following two reasons:

- using an SGML format to encapsulate the individual fields would guarantee a high level of compatibility with existing TREC collections, and
- a similar representation format as used with SAC would offer a smooth transition for our participants from the previously used SAC to the *IAPR TC-12 Image Benchmark* (e.g. to keep changes in existing retrieval scripts to a minimum).

Representation Language

Unlike in previous years where only English representations were offered to participants, we provided an additional language to the participants of *ImageCLEFphoto 2006* by offering a German version of the representations as well. Figure 7.3 displays

```

<DOC>
<DOCNO>annotations/16/16019.ger</DOCNO>
<TITLE>Der Flamingostrand</TITLE>
<DESCRIPTION> ein Photo eines braunen Sandstrands; das dunkelblaue
Meer mit kleinen brechenden Wellen dahinter; eine dunkelgrüne
Palme im Vordergrund links; ein blauer Himmel mit Wolken am
Horizont im Hintergrund; </DESCRIPTION>
<NOTES> Originalname auf portugiesisch: "Praia do Flamengo"; Der
Flamingostrand gilt als einer der schönsten Strände Brasiliens;
</NOTES>
<LOCATION>Salvador, Brasilien</LOCATION>
<DATE>2 Oktober 2002</DATE>
<IMAGE>images/16/16019.jpg</IMAGE>
<THUMBNAIL>thumbnails/16/16019.jpg</THUMBNAIL>
</DOC>

```



Figure 7.3: The generated German caption for image 16019.jpg.

an example of the German representation for the image 16019.jpg, whereby only the content of the tags is translated, while the tags themselves remain in their original English versions. Having two sets of representations in different languages is beneficial for a multilingual evaluation environment such as *ImageCLEF* as it allows for the creation of many interesting retrieval and evaluation scenarios, including:

- the comparison of English and German monolingual retrieval;
- the comparison of translation directions (*i.e.* does English retrieval from German documents perform better than German retrieval from English documents?);
- the evaluation of translation resources for third languages (*e.g.* the comparison of retrieval performance based on Spanish-to-English against Spanish-to-German translations);
- the investigation whether combined retrieval from both collections would outperform the results based on monolingual retrieval.

We did not offer the Spanish versions of the logical image representations as they were still being verified and were not in a release status yet.

Representation Completeness

Since consistent and careful semantic descriptions of images are typically not found in practice, we decided to create a more realistic scenario for participants by releasing a subset of the collection with a varying degree of representation “completeness” (*i.e.* with different representation fields available for indexing and retrieval). Thus for *ImageCLEFphoto 2006*, we generated a subset that covered the following levels of completeness:

- 70% of the semantic descriptions contained title, description, notes, location and date;
- 10% of the semantic descriptions contained title, location and date;
- 10% of the semantic descriptions contained location and date; and
- 10% of the images were not annotated (or had empty tags respectively).

This distribution of representation completeness would allow for the subsequent analysis of whether more visual approaches would improve the retrieval results for topics that predominately target images with incomplete textual representations.

Orthography

We did not make use of the possibility of an additional orthographic challenge by further injecting spelling mistakes or typographical errors into the logical image representations. Although one might argue that this would reflect realistic data found in generic photographic collections (especially in private ones), the main goal of *ImageCLEFphoto 2006* was to evaluate systems by their ability to retrieve relevant images, and not by the ability of detecting and correcting misspelled words. We therefore set the level of orthography to 100% (*i.e.* no typographical errors introduced) during the generation of the text files.

7.2.3 Topic Release

We gave the participants six weeks to familiarise themselves with the new collection so that they could (1) adapt their existing retrieval scripts to the new multilingual image representations and/or (2) extract the visual and textual features of the images and their logical representations in order to index the entire collection. In the meantime, we had created 60 topics (see Table 5.9 in Section 5.4.2) representing typical search requests for the *IAPR TC-12 image collection* and finally released them to the participants on 15 April 2006.

Topic Components and Format

Each original topic comprised a title (a short sentence or phrase describing the search request in a few words), a narrative (a description of what constitutes a relevant or non-relevant image for each request), and three image examples (these images were not removed from the collection, but removed from the set of relevance judgments). Figure 7.4 displays an example for a generated English topic as given

```
<top>
<num> Number: 14 </num>
<title> scenes of footballers in action </title>
<narr> Relevant images will show football (soccer)
players in a game situation during a match. Images with
footballers that are not playing (e.g. players posing for
a group photo, warming up before the game, celebrating
after a game, sitting on the bench, and during the half-
time break) are not relevant. Images with people not
playing football (soccer) but a different code (American
Football, Australian Football, Rugby Union, Rugby League,
Gaelic Football, Canadian Football, International Rules
Football, etc.) or some other sport are not relevant.
</narr>
<image> images/31/31609.jpg </image>
<image> images/31/31673.jpg </image>
<image> images/32/32467.jpg </image>
</top>
```



Figure 7.4: Topic with three sample images.

to the participants. The format of the topic file was identical with the one used in previous years: the entire topic was encapsulated by the `<top>` and `</top>` tags, the `<num>` tag uniquely identified the topic (numbers from 1 to 60), while the topic title and the narrative description were embedded in the `<title>` and `<narr>` tags respectively. Further, for *ImageCLEFphoto 2006*, we also introduced the new tag `<image>`, which contained the path to the sample images that could subsequently be used for visual based approaches.

Topic Translation

Retrieval from generic photographic collections such as the *IAPR TC-12 image collection* is inherently multilingual, thus one key part of evaluation in *ImageCLEFphoto* was to provide queries in a language different from that used to describe the images.

As a consequence, we translated the topic titles into 15 languages: German, Spanish, French, Italian, Portuguese, Dutch, Russian, Polish, Danish, Swedish, Finnish, Norwegian, Japanese, and Simplified and Traditional Chinese. The choice of languages was based on previous submissions to *ImageCLEF* (these 15 languages were exactly the ones that had actually been used in *ImageCLEF 2005*) and on the feedback of the participants.

All translations were provided by at least one native speaker and verified by at least another native speaker. Unlike in past campaigns, however, we did neither translate nor evaluate the topic narratives, because they were only created to unambiguously define what constitutes relevant and non-relevant images for each topic and did not present a realistic search scenario (users are not very likely to enter such a long query into a concept-based search engine).

Visual Topics

To further investigate the success of visual techniques, thirty topics from *ImageCLEFphoto 2006* were selected and modified to reduce semantic information and make better suited to visual retrieval techniques. For example, removing geographic

constraints (e.g. “black and white photos” instead of “black and white photos *from Russia*”) and other, non-visual constraints (e.g. “*child* wearing baseball cap” instead of “*godson* wearing baseball cap”). Table 7.1 displays the title of the visual topics.

ID	Topic Title	ID	Topic Title
61	church with more than two towers	76	people on surfboards
62	group in front of mountain landscape	77	group pictures on a beach
63	animal swimming	78	bird flying
64	straight road	79	photos with Machu Picchu in background
65	group standing in salt pan	80	Machu Picchu and Huayna Picchu in bad weather
66	black and white photos	81	winter landscape
67	scenes of footballers in action	82	sunset over water
68	night shots of cathedrals	83	images of typical Australian animals
69	lighthouses at the sea	84	indoor photos of churches or cathedrals
70	close-up photograph of an animal	85	photos of dark-skinned girls
71	tennis player on tennis court	86	views of walls with asymmetric stones
72	snowcapped buildings	87	television and telecommunication towers
73	child wearing baseball cap	88	drawings in deserts
74	motorcyclists riding on racing track	89	photos of oxidised vehicles
75	exterior view of churches or cathedrals	90	salt heaps in salt pan

Table 7.1: The visual topics.

We wanted to attract more visually orientated groups to *ImageCLEFphoto* which to date has been dominated by groups using textual approaches. Participants were given three example images to describe each topic and were required to perform query-by-visual-example retrieval to begin the search. To strictly separate these additional visual topics from the “official” topic set, we assigned them identifiers between 61 and 90.

The 30 visual topics were further classified into three evenly sized groups according to how visual they were estimated to be (the same approach as described in the *Visual Retrieval Challenges* paragraph of Section 5.4.2). Based on these findings, the topics were categorised into 10 *easy* topics that should do well with CBIR techniques (level > 3), 10 *hard* topics that will be quite difficult for CBIR (level ≤ 2), and 10 *medium* topics that should lie in between these two categories ($2 < \text{level} \leq 3$). The exact distribution of the topics across this dimension can be found in Appendix A.

7.2.4 Submission of Runs

We gave the participants six weeks to perform their retrieval experiments and to submit their results (*runs*) by 1 June 2006. The participants were allowed to submit as many runs as they wished for to investigate different approaches; out of the 36 groups that had registered for *ImageCLEFphoto 2006*, 12 eventually submitted a total of 157 runs.

Submission Format

The participants were required to submit a ranked list of (up to) 1000 images for each of the topics, with images being ranked in descending order of similarity: the higher the rank of an image, the more likely it is to be relevant. The submissions for *ImageCLEFphoto 2006* thereby followed the standard TREC format, which requires participants to submit a text file organised in six columns:

1. The first column is the topic number (1-60 in 2006).
2. The second column is the query number within that topic which should allow for variations between the translations (not used in *ImageCLEFphoto 2006*).
3. The third column is the official document number of the retrieved image in the form of: directory/filename, *e.g.* 15/15001, where the filename has the extension removed.
4. The fourth column is rank position.
5. The fifth column shows the score (integer or floating point) that generated the ranking in descending order.
6. The sixth column contains the *run tag*, a unique identifier for each group and method used.

A detailed description of these columns and several examples can be found on the web page² of *ImageCLEFphoto 2006*.

²<http://eurovision.shef.ac.uk/~cloughie/cgi-bin/imageclef2006/adhoc.htm>

Submission Guidelines

The participants were free to experiment with whatever methods they wished for CLIR, TBIR and CBIR. Examples include query expansion based on thesauri or relevance feedback, different models of retrieval, different translation resources (*e.g.* dictionary-based vs. machine-translation), and combining concept-based and content-based methods for retrieval. To enable participation for research groups without access to their own CBIR system, we provided access to GIFT and FIRE (see Section 2.7.5). We further asked the participants to submit a monolingual baseline run (English-English or German-German) which could subsequently be used to evaluate the translation performance of bilingual runs.

Submission Categorisation

Rather than listing all the different possible approaches that could be used to perform retrieval, we asked the participants to categorise their submissions according to the following dimensions:

- query language (any of the 16 languages offered);
- annotation language (English, German or both);
- run type (automatic or manual);
- use of feedback or automatic query expansion;
- modality (text only, image only or combined).

Participants had to indicate these dimensions in their run identifiers to allow for the subsequent comparison with other submissions. For example, the baseline run for English-English would have been identified as `EN-EN-AUTO-NOFB-TXT`.

7.2.5 Relevance Assessments

As soon as we had received all the submissions, we (*i.e.* the two topic creators) started to carry out the relevance assessments using the third component of the benchmark administration system: the module for relevance judgments.

Standard Topics

We decided on the *pooling method* and used the top 40 results from all submitted runs (for the topics 1 - 60) to create image pools giving an average of 1,045 images to judge per topic. Figure 7.5 provides an overview of the pool sizes for each topic

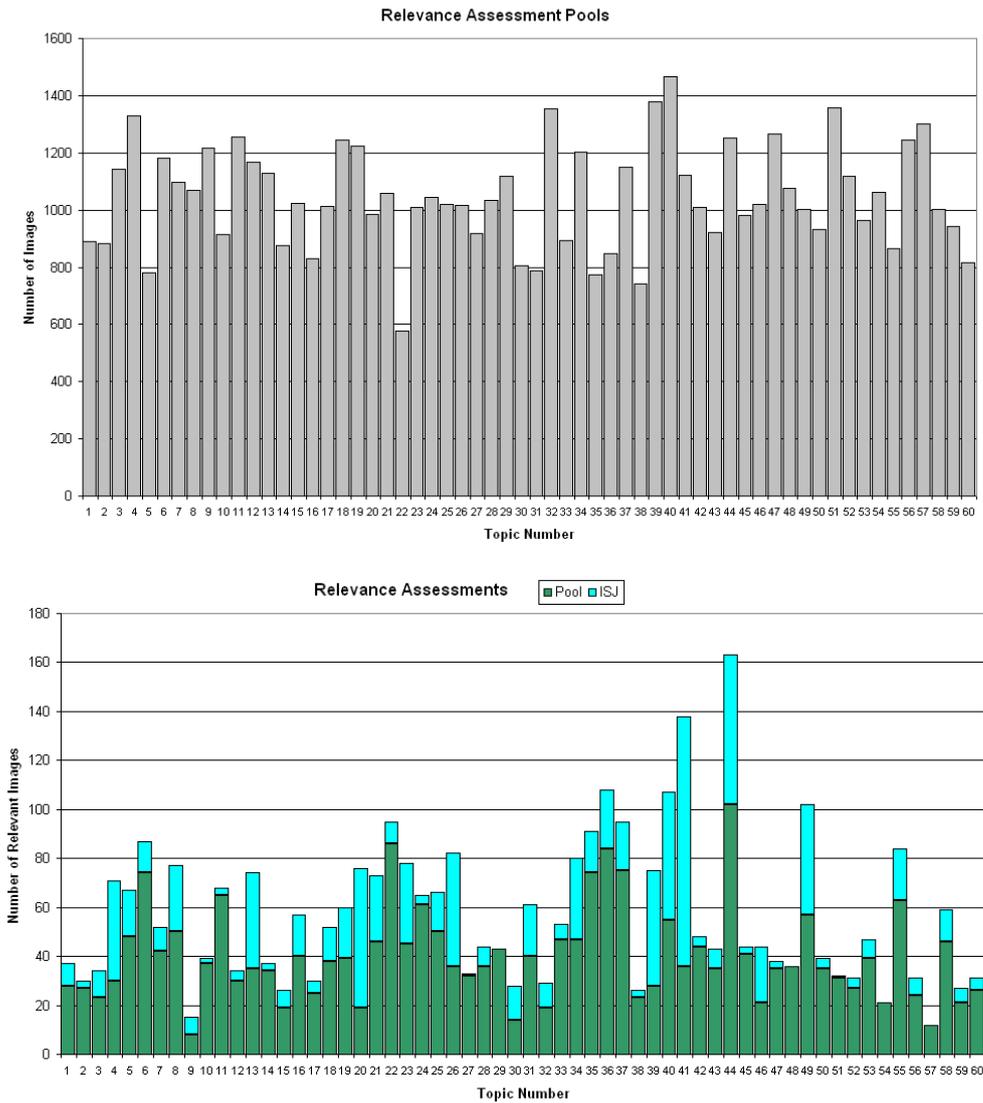


Figure 7.5: Relevance assessments (ad-hoc topics).

(above) and the number of relevant images found in the pools as well as those added through the use of ISJ (below). We judged all the images in the topic pools and also used ISJ to supplement the pools with further relevant images; on average,

25.24% of the relevant images were added using ISJ. Table 7.2 provides a statistical

Relevance assessments	Average	Minimum	Median	Maximum	σ
Pool size (images)	1044.9	575	1022	1468	182.3
Total relevant (images)	57.1	12	50	163	30.0
Relevant through ISJ (in %)	25.2	0.0	22.1	75.0	18.8

Table 7.2: Relevance assessment statistics (ad-hoc topics).

overview of the number of images in the relevance pools, the number of relevant images and the percentage of additional relevant images found through ISJ.

Visual Topics

Figure 7.6 provides an overview of the pool sizes for each visual topic and the number of relevant images found in the pools as well as those added through the use of ISJ. The same methodology was applied for the relevance judgments of the visual topics (ID from 61 to 90), where we also used the top 40 results from all submitted runs and created image pools giving an average of 171 images to judge per topic. The rather small pool sizes combined with the rather weak retrieval results led to heavy use of ISJ to complement the pools with further relevant images; on average, 77.45% of relevant images were added by ISJ. Table 7.3 provides a statistical overview of

Relevance assessments	Average	Minimum	Median	Maximum	σ
Pool size (images)	170.8	83	176	196	23.9
Total relevant (images)	100.3	22	69.5	419	100.0
Relevant through ISJ (in %)	77.5	31.6	83.6	98.0	18.2

Table 7.3: Relevance assessment statistics (visual topics).

the number of images in the relevance pools, the number of relevant images and the percentage of additional relevant images found through ISJ for the visual topics.

Assessment Methodology

Although the very detailed narrative descriptions had clearly defined what constitutes a relevant image, we based our judgments on a ternary classification scheme to deal with any potential uncertainties during the assessment: images were either

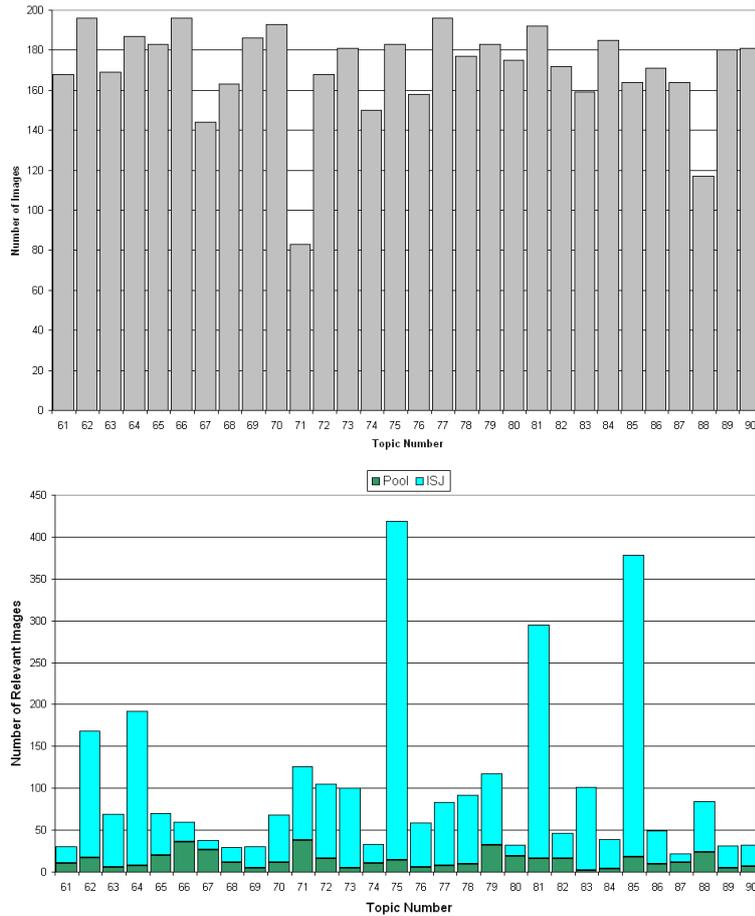


Figure 7.6: Relevance assessments (visual topics).

(1) relevant, (2) partially relevant or (3) not relevant. Based on these judgments, we only considered those images for the set of relevant images (qrels) which had been judged as relevant by both assessors (*intersect-strict*).

The ISJ was based (1) on textual searches, (2) on the topic creators' profound knowledge of and familiarity with the collection, and (3) on the predefined ground-truth that we had established to estimate the size of the target set in the topic creation process (compare Section 5.4).

Appendix A provides more information on the exact pool sizes and the percentage of images added using ISJ for each individual topic of both the standard ad-hoc set (1 - 60) and the additional visual set (61 - 90) of topics.

7.2.6 Result Generation and Notification

Once these relevance judgments were completed, we were able to evaluate the performance of the individual systems and approaches (with the deadline for this result generation process being 15 July 2006).

We computed the results for the submitted runs using the latest version of `trec_eval`³, which provided us with over 130 performance measures for each system run, and we decided to rank the retrieval performance of the submitted runs according to the following measures: *MAP* as the primary measure, and *P(20)*, *GMAP* and *bpref* as additional measures (see below). We provided the participants with individual rankings for each of the measures (as well as with one combined ranking which treated each of the four measures equally by simply averaging the ranks of all four measures for each system).

Mean Average Precision

Following the TREC-style tradition, the primary measure for system evaluation was the un-interpolated (arithmetic) *mean average precision (MAP)*, which is currently one of the leading performance measures in many ad-hoc retrieval evaluation events because it is a very stable measure with a low error rate and is based on an abundance of information (*e.g.* it represents the area underneath the highly informative precision-recall graph).

Further, according to participants' feedback in the 2004 and 2005 campaigns, the general consensus between researchers was that, although there are also some cons to using MAP, it should be kept as the primary measure for the evaluation event because it rewards an algorithm's ability to rank relevant images which, in fact, is the main goal given for the evaluation.

Precision at Rank 20

The ad-hoc retrieval task from generic photographic collections models the scenario that is also given in many online search engines such as *Google* or *Yahoo!*: find as

³http://trec.nist.gov/trec_eval/trec_eval.7.3.tar.gz

many relevant images as possible to a given statement of a user information need. This scenario is generally more concerned with precision than recall (users want to see relevant images on the first result page, while it is not of primary importance to them that all the relevant images are found on subsequent result pages), and since most of these search engines display 20 images on their first result page by default⁴, we decided to include $P(20)$ as one of our additional measures.

Geometric Mean Average Precision

ImageCLEFphoto is the first VIR evaluation event to include the *geometric mean average precision (GMAP)* as a performance measure - it had only been used in text retrieval tasks such as [471, 472] so far.

One goal of the evaluation was to observe and analyse retrieval effectiveness with respect to topic difficulty levels. However, this is often difficult using *MAP* and $P(20)$ as these measures allow better performing (easy) topics to mask changes in the scores of poorly performing (difficult) topics. We therefore introduced *GMAP* as an additional measure in order to highlight difficult topics, because it emphasises topic scores close to 0.0 (the “bad results”) while minimising differences between larger scores (the “good results”) and therefore does not let better performing topics mask weaker ones. It is, further, a very robust measure that remains highly stable with as few as 50 topics.

Binary Preference

Finally, we also considered the *binary preference (bpref)* as another additional measure, because *bpref* allows for some control over the quality of relevance assessments. This measure is a function of the number of times *judged* non-relevant images are ranked before relevant ones, and it is therefore also a good indicator for the completeness of relevance judgments [35].

⁴Google, Yahoo! and Altavista do so as of 31 March 2007.

7.2.7 Evaluation Event and Publication

Once we had provided the participants with the ranked lists of their system runs, they had about four weeks to create their preliminary workshop papers (in which they described their approaches and analysed their achieved retrieval performance) and to send them to CLEF by 15 August 2006.

The *Cross Language Evaluation Forum* then took place in Alicante, Spain from 21 to 23 September 2006. The participants met with the organisers to present their systems and to compare them on grounds of the evaluation results⁵. Moreover, in a special break-out session, we asked the participants for their feedback (see Section 7.4.4) and discussed potential future directions and evaluation tasks for *ImageCLEFphoto 2007* and onwards.

After CLEF, the participants had about two months until December 2006 to finalise their papers describing all the novel techniques, new findings and evaluation results. We then reviewed, revised and selected the papers that would eventually be printed in the *Springer* proceedings under the series of *Lecture Notes in Computer Science* (LNCS), which finally completed the evaluation event.

7.3 Retrieval Performance Analysis

While the previous section explained the methodology and illustrated the organisation and realisation of *ImageCLEFphoto 2006*, this section will now concentrate on the description of the retrieval systems and provide an analysis of their performance according to several submission parameters and topic dimensions.

7.3.1 Submission Overview

Out of the 36 groups that had registered for *ImageCLEFphoto 2006*, 12 also submitted a total of 157 runs (all of which were evaluated). Table 7.4 summarises the participating groups and the number of runs submitted by them. The 12 groups

⁵These results do not necessarily determine whether a paper is accepted for an oral presentation; other factors like originality of the paper, novelty of the technique and/or political reasons also come into play.

Group ID	Institution	Runs
Berkeley	University of California, Berkeley, USA	7
CEA-LIC2M	Fontenay aux Roses Cedex, France	5
CELI	CELI srl, Torino, Italy	9
CINDI	Concordia University, Montreal, Canada	3
DCU	Dublin City University, Dublin, Ireland	40
IPAL	IPAL, Singapore	9
NII	National Institute of Informatics, Tokyo, Japan	6
Miracle	Daedalus University, Madrid, Spain	30
NTU	National Taiwan University, Taipei, Taiwan	30
RWTH	RWTH Aachen University, Aachen, Germany	6
SINAI	University of Jaén, Jaén, Spain	12
TUC	Technische Universität Chemnitz, Germany	4

Table 7.4: Participating groups.

were from 10 different countries and 3 continents, again illustrating a very global and international field of participation. Each of the participants made use of the fact that they could submit more than one run and submitted a minimum of three runs, with three groups handing in even 30 or more runs.

Submissions by Dimensions

Table 7.5 provides an overview of all submitted runs according to several dimensions. Most submissions used the textual image representations, with eight groups

Dimension	Type	Runs	Groups
Query language	bilingual	93	8
	monolingual	57	11
	visual	7	3
Annotation language	English	133	11
	German	18	4
	none	6	2
Modality	Text only	108	11
	Text + Image	43	7
	Image only	6	2
Query expansion	without	85	11
	with	72	8

Table 7.5: Submission overview by dimensions.

submitting bilingual runs and 11 groups monolingual runs. A total of 11 groups provided text-only runs, and for seven groups (CEA, CINDI, DCU, IPAL, Miracle,

NTU and TUC), the main focus of their submission was on combining text and visual features. Moreover, eight groups (Berkeley, CINDI, DCU, IPAL, Miracle, NTU, SINAI and TUC) used query expansion techniques to further improve their retrieval results. Many groups (*e.g.* Berkeley, DCU, NII, NTU and SINAI) made use of machine translation (MT) systems to translate the topics.

Based on all submitted runs, 59% were bilingual, 31% involved the use of image retrieval (27% using combined visual and textual features), and 46% of runs made use of query expansion techniques. The majority of runs were automatic (*i.e.* involving no human intervention), with only one run submitted being manual.

Submission by Languages

Table 7.6 displays the number of groups and participants per query and caption language. All groups (with the exception of RWTH) submitted at least one mono-

Query language	Caption language	Runs	Groups
English	English	49	11
Italian	English	15	4
Japanese	English	10	4
Simplified Chinese	English	10	3
French	English	8	4
Russian	English	8	3
German	English	7	3
Spanish	English	7	3
Portuguese	English	7	3
Dutch	English	4	2
Traditional Chinese	English	4	1
Polish	English	3	1
Visual	English	1	1
German	German	8	4
English	German	6	3
French	German	3	1
Japanese	German	1	1
Visual	(none)	6	3
Visual topics	(none)	6	2

Table 7.6: Ad-hoc experiments listed by query and caption languages.

lingual English run, while four groups also submitted a total of eight monolingual German runs. The majority of runs was concerned with retrieval from English

image representations, while only 11% of the monolingual and 14% of the bilingual experiments made use of the German ones. Unlike in previous years, where many participants had investigated Spanish and French, the most popular query languages for bilingual retrieval in *ImageCLEFphoto 2006* were Italian (15 runs), French (11 runs), Japanese (11 runs) and Simplified Chinese (10 runs).

7.3.2 Participating Groups and Methods

This section provides a brief description of the methods used in the submitted runs of each group (listed alphabetically by their group identifier) to provide a snap-shot of current research interests as well as an overview of the many novel approaches investigated at *ImageCLEFphoto*.

Berkeley

The *School of Information Management and Systems* of the *University of California* in Berkeley, USA, submitted 7 runs. All runs were text only: 4 monolingual English, 2 monolingual German, and one bilingual English-German. Further, 3 runs used query expansion in the form of pseudo-relevance feedback, and another 3 runs made use of the topic title and narratives.

The retrieval algorithm used a form of logistic regression with blind relevance feedback (the 10 highest weighting terms from the top 10 documents). Translation using *Babelfish* and expanding queries using the meta-data of relevant images was found to work well. An interesting result was that using query expansion *without* any translation of terms worked surprisingly well for the bilingual run [225].

CEA-LIC2M

The *CEA-LIC2M* group from Fontenay aux Roses Cedex, France, submitted five runs without using feedback or query expansion. The group submitted 2 visual runs, 2 concept-based runs and one that combined textual and visual features. Two runs were monolingual English, and one run was bilingual French-to-English.

Separate initial queries were performed using the textual and visual components of the topics, and then merged *a posteriori*. Documents and queries were preprocessed using a linguistic analyser, and performing visual retrieval on each query image and merging results appeared to provide better results than visual retrieval using all three example images simultaneously [28].

CELI

The participants from *CELI srl* of Torino, Italy, submitted 9 text-only, automatic runs (1 monolingual English and 8 bilingual Italian-English), whereby 6 of them made use of different query expansion techniques.

Translation was achieved using bilingual dictionaries, and a disambiguation approach based on latent semantic analysis was implemented. Using a boolean “AND” operator of the translations was found to provide higher results than using an “OR” operator. Further, the use of query expansion was shown to increase retrieval effectiveness to bridge the gap between the uncontrolled language of the query and the controlled language of the image meta-data [76].

CINDI

The CINDI group from *Concordia University* in Montreal, Canada, submitted 3 monolingual English runs: 2 text only and 1 mixed, 2 automatic and 1 manual, 2 with feedback (and 1 without), and 2 with query expansion (and 1 without).

The use of manual relevance feedback and the integration of text and image achieved the best performance for this group [350].

DCU

Dublin City University from Dublin, Ireland, submitted a total of 40 automatic runs, whereby 26 were text-only and 14 of mixed modality, and 27 with feedback and 13 without. DCU submitted 6 monolingual and 34 bilingual runs and explored 10 different query languages as well as both representation languages.

Concept-based retrieval was performed using the *BM25* weighting operator, and

visual features were matched using the JD. Image retrieval on individual images was performed and merged using the *CombMAX* operator, while textual and visual runs were fused using the weighted *CombSUM* operator. The results showed that fused text and image retrieval consistently outperformed text-only methods, and that the use of pseudo relevance feedback improved the effectiveness of the concept-based retrieval model [93].

IPAL

IPAL, Singapore, submitted 13 automatic, monolingual runs (6 visual, 4 mixed and 3 text only) and a further 4 runs to the visual-only subtask.

They tested various indexing methods, used the *XIOTA* system for text retrieval and also applied pseudo relevance feedback. For the visual topics, the query images and all the images of the collection were indexed with feature reduction using LSI, and the images were then ranked according to their distances to the query images. Their results indicate, again, that the combination of text and image retrieval leads to better performance [258].

Miracle

The *Miracle group* of the *Daedalus University* in Madrid, Spain, submitted 30 automatic runs (28 text only, 2 mixed) and a further 10 runs involving query expansion based on *WordNet*. The group only used the English image representations and generated 18 monolingual English and 12 bilingual runs (Russian, Polish, Japanese and simplified Chinese). A total of 8 runs used narrative descriptions only, 9 runs used both title and narratives, and the remaining 11 runs used the titles only.

The most effective approach was shown to be the indexing of nouns from the logical image representations with no other processing [269].

NII

The *National Institute of Informatics* from Tokyo, Japan, submitted 6 text-only automatic runs without feedback or query expansion, exploiting all possibilities

of the three languages English, German and Japanese: 1 monolingual English, 1 monolingual German and 4 bilingual runs.

NII used the *Lemur* toolkit for text retrieval, *Babelfish* for translation and experimented with a visual feature-based micro-clustering algorithm to link nearly identical images annotated in different languages. This clustering approach, however, did not improve retrieval effectiveness [184].

NTU

The *National Taiwan University* from Taipei, Taiwan, also submitted 30 automatic runs: 10 text only and 20 mixed, and 12 with and 18 without feedback. Further, a total of 2 monolingual English and 2 monolingual German runs, 1 visual run and 25 bilingual runs (using English image representations only) exploring 10 different languages were handed in.

NTU showed that the use of visual features could improve text-only retrieval based on the logical image representations. A novel word-image ontology approach did not perform as well as retrieval using the image representations alone. *Systran* was used to provide translation, and the initial query images were found to improve ad-hoc retrieval [53].

RWTH

The *Human Language Technology and Pattern Recognition Group* of *RWTH Aachen University* from Aachen, Germany, submitted a total number of 4 entirely visual runs: 2 for the standard ad-hoc task, and 2 for the visual retrieval sub-task.

Two different approaches were attempted in both tasks: one approach saw the use of invariant and Tamura texture feature histograms, which were compared using JSD, weighing IFH twice as strong as texture features based on the assumption that colour information outranks texture information for databases of general photographs; in the other approach, they used 2048 bin histograms of image patches in colour, which were compared according to their colour and texture using JSD. Visual-only retrieval did not perform well in either task [88].

SINAI

The *SINAI* group of the *University of Jaén*, Spain, submitted 12 automatic text-only runs (8 runs with query expansion) using only the English image representations. The group submitted 4 monolingual runs and 8 bilingual runs using the Dutch, French, German, Italian, Portuguese and Spanish topics.

They used a number of different machine translation (MT) systems to translate these topics and the *Lemur* toolkit implementation of *Okapi* as the underlying retrieval model. Their results indicate that retrieval based on simple probabilistic models such as *td-idf* or *Okapi* is not very effective for concept-based image retrieval unless pseudo-relevance feedback techniques are applied [89].

TUC

Technische Universität Chemnitz from Chemnitz, Germany, submitted four automatic monolingual English runs: 3 text only and 1 mixed as well as 3 with feedback (and query expansion) and 1 without.

Combining and/or merging independent content and concept-based runs appeared to give the highest retrieval effectiveness, together with the use of concept-based query expansion [489].

7.3.3 System Performance Analysis

The absolute retrieval results achieved by the systems were lower in 2006 compared to previous years. We attribute this to the choice (and increased difficulty) of topics, a more visually challenging photographic collection and there being incomplete semantic representations provided with the collection. This section provides an overview of the system results with respect to query and representation languages as well as other submission dimensions such as query mode, retrieval modality and the involvement of relevance feedback or query expansion techniques.

Results by Language

Table 7.7 shows the runs which achieved the highest MAP for each language pair (ranked by descending order of MAP scores). Of these runs, 83% used feedback

Language (Captions)	Group	Run ID	MAP	P(20)	BPREF	GMAP
English (English)	CINDI	Cindi_Exp_RF	0.385	0.530	0.874	0.282
German (German)	NTU	DE-DE-AUTO-FB-TXTIMG	0.311	0.335	0.974	0.132
Portuguese (English)	NTU	PT-EN-AUTO-FB-TXTIMG	0.285	0.403	0.755	0.177
T. Chinese (English)	NTU	ZHS-EN-AUTO-FB-TXTIMG	0.279	0.464	0.669	0.154
Russian (English)	NTU	RU-EN-AUTO-FB-TXTIMG	0.279	0.408	0.755	0.153
Spanish (English)	NTU	SP-EN-AUTO-FB-TXTIMG	0.278	0.407	0.757	0.175
French (English)	NTU	FR-EN-AUTO-FB-TXTIMG	0.276	0.416	0.750	0.158
Visual (English)	NTU	AUTO-FB-TXTIMG	0.276	0.448	0.657	0.107
S. Chinese (English)	NTU	ZHS-EN-AUTO-FB-TXTIMG	0.272	0.392	0.750	0.168
Japanese (English)	NTU	JA-EN-AUTO-FB-TXTIMG	0.271	0.402	0.746	0.170
Italian (English)	NTU	IT-EN-AUTO-FB-TXTIMG	0.262	0.398	0.722	0.143
German (English)	DCU	combTextVisual.DEENEN	0.189	0.258	0.683	0.070
Dutch (English)	DCU	combTextVisual.NLENEN	0.184	0.234	0.640	0.063
English (German)	DCU	combTextVisual.ENDEEN	0.122	0.175	0.524	0.036
Polish (English)	Miracle	miratctdplen	0.108	0.139	0.428	0.005
French (German)	DCU	combTextVisual.FRDEEN	0.104	0.147	0.245	0.002
Visual (none)	RWTH	RWTHi6-IFHTAM	0.063	0.182	0.366	0.022
Japanese (German)	NII	mcp.bl.jpntger_td.skldir	0.032	0.051	0.172	0.001

Table 7.7: Systems with highest MAP for each query language.

of some kind (typically pseudo relevance feedback), and a similar proportion used both visual and textual features for retrieval. It is interesting to note that English monolingual runs outperforms the German monolingual ones (19% lower), and that German retrieval from English image representations produces better results than English retrieval from German collections (35% lower).

Further, the highest bilingual to English run was Portuguese to English, which performed 74% of the monolingual results, but the highest bilingual to German run was English to German which performed only at only 39% of the monolingual results. Also, unlike in previous years, the top-performing bilingual runs have involved Portuguese, traditional Chinese and Russian as the source language, showing an improvement of the retrieval methods using these languages.

Results by Query Mode

Table 7.8 illustrates the average scores across all systems runs (and the standard deviations in parenthesis) with respect to monolingual, bilingual and purely visual retrieval. While visual methods alone showed a rather weak retrieval performance

Query mode	MAP	P(20)	BPREF	GMAP
Monolingual	0.1538 (0.0898)	0.2141 (0.1183)	0.5157 (0.2289)	0.0594 (0.0671)
Bilingual	0.1443 (0.0735)	0.1946 (0.1049)	0.5168 (0.2151)	0.0379 (0.0442)
Visual	0.0743 (0.0900)	0.1788 (0.1224)	0.3475 (0.1482)	0.0272 (0.0357)

Table 7.8: Results by query mode.

(as expected), it was quite interesting to notice that bilingual retrieval only performed slightly lower than monolingual, which indicates that translation resources have advanced and can provide automatic translation on a high satisfactory level. On the other hand, we also attribute this to the fact that the nature of this task was, in general, a VIR challenge, whereby topic translation only constitutes one out of many dimensions and might hence not have had a major influence on the retrieval results.

Results by Representation Language

Table 7.9 illustrates the average scores across all systems runs (and the standard deviations in parenthesis) with respect to the representation languages: on average, MAP results for English as the target language are 26% higher than those for German (the statistic is significant at the 0.05 level using the *Student's t-test*). Reasons

Language	MAP	P(20)	BPREF	GMAP
English	0.1515 (0.0820)	0.2069 (0.1149)	0.5271 (0.2234)	0.0496 (0.0564)
German	0.1213 (0.0698)	0.1662 (0.0847)	0.4380 (0.2248)	0.0206 (0.0298)
None	0.0408 (0.0159)	0.1340 (0.0338)	0.2959 (0.0630)	0.0139 (0.0056)

Table 7.9: Results by image representation language.

for these findings might include better translation resources for bilingual retrieval from English collections, the larger variety of more sophisticated query processing techniques (stemmers, lemmatisers, query expansion) optimised for English than for German (the more complex grammar of German makes stemming an additional challenge as well), or simply the fact that more participants investigated retrieval from English than from German collections (as the latter one had only been introduced in 2006 and constituted a novel retrieval challenge for our participants).

Results by Retrieval Modality

In previous years, the system results had shown that combining visual features from the image and semantic knowledge derived from the logical image representations had offered optimum performance for retrieval from a collection of historic photographs. As indicated in Table 7.10, the results of *ImageCLEFphoto 2006* show

Modality	MAP	P(20)	BPREF	GMAP
Combined	0.1988 (0.0772)	0.2814 (0.1144)	0.6496 (0.1482)	0.0949 (0.0650)
Text Only	0.1288 (0.0619)	0.1727 (0.0800)	0.4646 (0.1661)	0.0272 (0.0365)
Image Only	0.0408 (0.0159)	0.1340 (0.0338)	0.2959 (0.0630)	0.0139 (0.0056)

Table 7.10: Results by retrieval modality.

that this also applies for retrieval from general collections of generic photographs: on average, combining visual features from the image and semantic information from the logical image representations gave a 54% improvement over retrieval based solely on text.

Results by Feedback and/or Query Expansion

The use of query expansion was shown to increase retrieval effectiveness by bridging the gap between the languages of the query and the logical image representations. In general, feedback (typically in the form of query expansion based on pseudo relevance feedback) also appears to work well on short image representations and is likely due to the limited vocabulary exhibited by these semantic descriptions. As

Feedback	MAP	P(20)	BPREF	GMAP
With	0.1646 (0.0900)	0.2239 (0.1278)	0.5475 (0.2082)	0.0622 (0.0670)
Without	0.1277 (0.0548)	0.1816 (0.0693)	0.4761 (0.1525)	0.0309 (0.0363)

Table 7.11: Results by feedback or query expansion.

displayed in Table 7.11, using some kind of query expansion or feedback (visual and textual) gives a 39% improvement over runs without it. Combined media and feedback runs had performed the highest for the evaluation of retrieval from historic photographs in previous years, a trend which now has been verified for retrieval from generic collections as well.

Visual Topics

Most runs submitted to the visual sub-task (as displayed in Table 7.12) showed quite promising results for precision values at a low cut-off rank, for example $P(20) = 0.285$ for the best run. However, it is felt that this is because some relevant images

RK	RUN ID	MAP	P(20)	BPREF	GMAP
1	RWTHi6-IFHTAM	0.1010	0.2850	0.4307	0.0453
2	RWTHi6-PatchHisto	0.0706	0.2217	0.3831	0.0317
3	IPAL-LSA3-VisualTopics	0.0596	0.1717	0.3360	0.0281
4	IPAL-LSA2-VisualTopics	0.0501	0.1800	0.3093	0.0218
5	IPAL-LSA1-VisualTopics	0.0501	0.1650	0.3123	0.0236
6	IPAL-MF-VisualTopics	0.0291	0.1417	0.2374	0.0119

Table 7.12: The visual results.

in the database are visually very similar to the query images, rather than the algorithms really understanding what is being searched for. The retrieved images at higher ranks appeared random, and further relevant images were only found by chance. This is also reflected by the low MAP scores (0.101 for the best run).

Image retrieval systems can, by all means, achieve decent results in retrieval tasks for specific domains, or in those that are well-suited to the current level of CBIR. However, the low results of the visual sub-task highlight the fact that the successful application of visual techniques in systems involving more general (and less domain-specific) pictures still requires much investigation.

7.3.4 Topic Analysis

There are considerable differences between the retrieval effectiveness of individual topics. For example, “photos of radio telescopes” has an average MAP of 0.5161, whereas “tourist accommodation near Lake Titicaca” has an average MAP of 0.0027. Possible causes for these different scores include:

- the discriminating power of query terms in the collection;
- the complexity of topics (*e.g.* the topic “Tourist accommodation near Lake Titicaca” involves a location and fuzzy spatial operator which will not be

handled appropriately unless support for spatial queries is provided);

- the level of semantic knowledge required to retrieve relevant images (this will limit the success of purely visual approaches); and
- the translation success for bilingual runs (*e.g.* whether proper names have been successfully handled).

We can further identify the following trends of MAP and $P(20)$ with respect to: submission modality, log file analysis, geographic constraints, visual features, topic difficulty and representation completeness.

Submission Modality

Figure 7.7 displays the average MAP across (all) system runs for each topic based on modality. Many topics clearly show an improvement through the use of combining

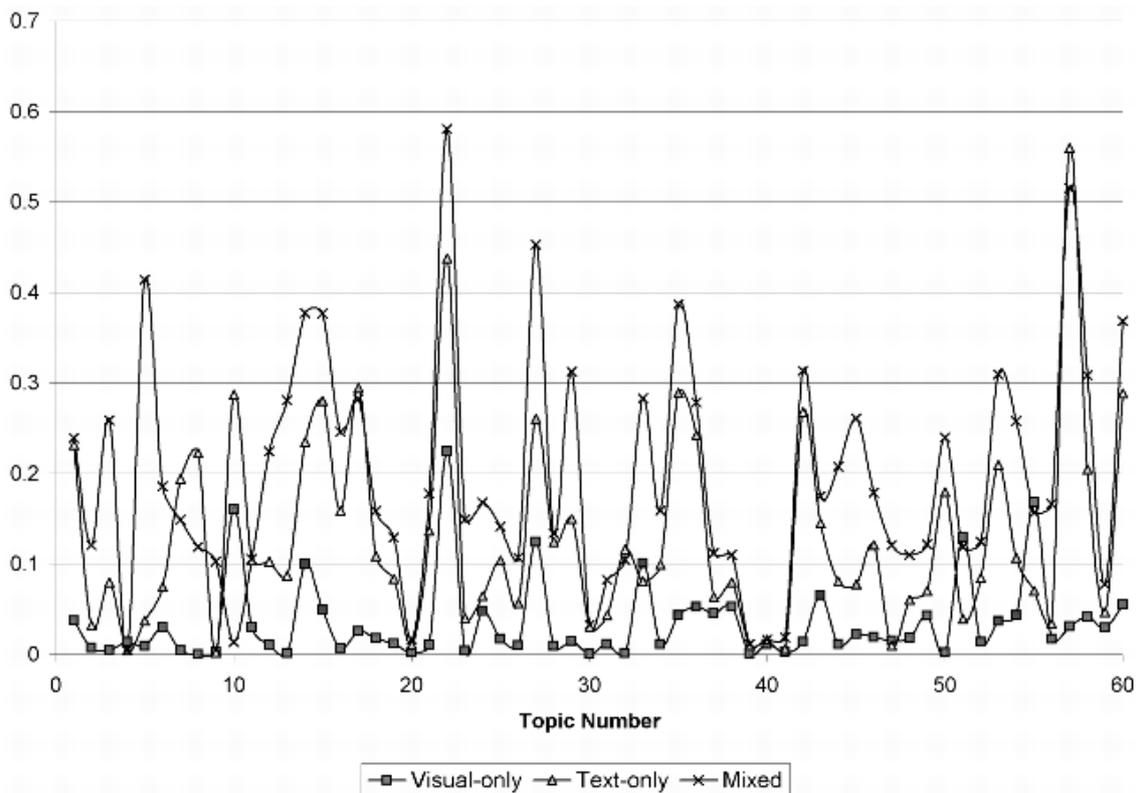


Figure 7.7: MAP by topic based on modality.

textual and visual features (mixed) than any single modality alone. Part of this is

likely to be attributed to the availability of visual examples with the topics which could be used in the mixed runs (and to the fact that these examples were directly taken from the collection).

Topic Origin

Table 7.13 displays the average retrieval performance according to $P(20)$ and MAP (with the standard deviations in parenthesis) for all topics with respect to their origin. Topics that were directly taken or derived from the log file thereby only

Topics ...	avg MAP	avg P(20)
directly taken from the log file	0.1296 (0.0928)	0.1987 (0.1335)
derived from the log file	0.1155 (0.0625)	0.1578 (0.0716)
not taken from the logfile	0.2191 (0.1604)	0.2172 (0.1063)

Table 7.13: Topic overview by topic origin.

achieved about 55% of the retrieval performance compared to those not taken from the log file, with the ones derived from the log file showing the lowest results. It is likely that most topics not derived from the log file were more “visual” and perhaps therefore simpler to execute, while those derived from the log file were often altered to include additional text-retrieval challenges (such as vocabulary mismatches or the use of abbreviations) and hence more difficult.

Geographic Constraints

Table 7.14 shows that topics specifying spatial operators and/or specific locations were outperformed by topics that included general locations, man-made objects or no geography at all. These results are not surprising because most groups did

Topics with...	avg MAP	avg P(20)
specific locations and spatial operators	0.1146 (0.0872)	0.1559 (0.0981)
general locations or manmade objects	0.1785 (0.1111)	0.2388 (0.1120)
no geography	0.1313 (0.1219)	0.1986 (0.1476)

Table 7.14: Topic overview by geographic constraints.

not use geographic information retrieval (GIR) methods, with especially the low

retrieval performance for geographic topics indicating that the involvement of such methods could potentially contribute to improve the retrieval precision of VIR systems.

Visual Features

Table 7.15 displays the retrieval results of topics categorised by the visual retrieval challenge they offer (see Section 5.4.2). We had expected that more visual topics

Topics where additional use of CBIR...	avg MAP	avg P(20)
will not improve results (levels 1 and 2)	0.1179 (0.1041)	0.1583 (0.0918)
might or might not improve results (level 3)	0.1318 (0.0940)	0.1933 (0.1272)
should improve results (levels 4 and 5)	0.2250 (0.1094)	0.3081 (0.1256)

Table 7.15: Topic overview by visual features.

(*e.g.* “sunset over water” is more visual than “pictures of female guides”, which one could consider more semantic) were likely to perform better given that many participants had made use of combined visual and textual approaches; and indeed, topics in categories indicating that visual techniques would not improve results (levels 1 and 2) or could possibly improve them (level 3) did, on average, only show 52% and 59% of the results (MAP) achieved by topics from categories indicating that visual techniques were expected to improve results (levels 4 and 5).

Topic Difficulty

Table 7.16 displays the retrieval results of topics categorised by the level of their estimated retrieval difficulty based on their linguistic complexity as well as the statistical relationship of topic elements with the document collection and a set of relevant documents (see Section 5.3.3 for the exact definition of these levels). Since

Topics rated as...	difficulty (d)	avg MAP	avg P(20)
(very) easy	$0 \leq d < 2$	0.4148 (0.1427)	0.4528 (0.1121)
medium	$2 \leq d < 3$	0.1742 (0.0875)	0.2454 (0.1215)
hard	$3 \leq d < 4$	0.1128 (0.0732)	0.1815 (0.1002)
very hard	$d \geq 4$	0.0196 (0.0138)	0.0603 (0.0532)

Table 7.16: Topic overview by difficulty level.

we had established a novel measure to quantify topic difficulty measure (see Sections 5.2 and 5.3) showing a strong negative correlation between topic difficulty and retrieval results, it was not surprising that the “hard” topics were clearly outperformed by the “easier” ones.

Representation Completeness

Table 7.17 displays the retrieval results of topics categorised by the level of representation completeness of their corresponding relevant images, and its results show that retrieval performance is not necessarily correlated with the completeness of the logical image representations. The use of non-text approaches is the likely cause

Topics with ... having complete representations	avg MAP	avg P(20)
all relevant images (100%)	0.1668 (0.1356)	0.1781 (0.0995)
80 - 99% of relevant images	0.1290 (0.0653)	0.2003 (0.0960)
60 - 79% of relevant images	0.1353 (0.1002)	0.2275 (0.1449)
less than 60% of relevant images	0.1198 (0.1027)	0.1666 (0.1282)

Table 7.17: Topic overview by logical image representation completeness.

of successful retrieval for the topics with relevant images containing incomplete representations.

7.4 Event Analysis and Evaluation

ImageCLEFphoto 2006 has certainly made a massive contribution to evaluating and analysing the performance of VIR in general and of the participating VIR systems in particular. One vital aspect not covered so far, however, is the evaluation of the evaluation event itself.

Therefore in this section, we first quantify the quality of the *IAPR TC-12 Image Benchmark* and present an analysis of *ImageCLEFphoto 2006* evaluating the difficulty of the query topics as well as the choice of performance measures, and then report on the feedback we received from both participating and non-participating groups. Based on these comments, we will finally present an outlook to the future, including some ideas for the organisation of *ImageCLEFphoto 2007*.

7.4.1 Benchmark Validation

Prior to any further evaluation and analysis, it is crucial to evaluate the quality of the *IAPR TC-12 Image Benchmark* as an IR test collection itself. Hence, in the following, we will use the *stability method* [34, 372] to quantify:

- the confidence associated with the decision that one submitted retrieval run is better than another;
- the power of the test collection to discriminate among retrieval runs;
- the overall performance of the *IAPR TC-12 Image Benchmark* in comparison with other IR test collections.

Validation Method

The stability method, first introduced by Buckley in 2000 [34], is based on the comparison of each retrieval run (A) with every other submitted retrieval run (B) on a randomly selected subset of the query topics. The pair-wise comparisons for all retrieval runs are repeated multiple times, whereby each time a different (randomly selected) subset of the topics is used. Then, for each of these pairs, one counts how often the first run outperforms the second run (denoted by $|A > B|$), how often the second run outperforms the first one ($|A < B|$), and how often the two runs are regarded as equivalent ($|A \equiv B|$).

These comparisons are thereby carried out with respect to a particular performance measure and a given *fuzziness value* f . The fuzziness value is the percentage difference between scores such that, if the difference is smaller than the fuzziness value, the two scores are deemed equivalent. For instance, if the fuzziness value is 0.05, any scores within 5% of one another are counted as equal.

The definition of the *error rate* (or *minority rate* as referred to as by Sakai [372]) is based on the assumption that for each pair of runs, the correct comparison is given by the greater of the better-than ($|A > B|$) and worse-than ($|A < B|$) values, while the lesser of those two values is the number of times a test result is

misleading or in error. Buckley therefore defines the *error rate* (ER) as the total number of errors across all run pairs divided by the total number of comparisons:

$$ER = \frac{\sum \min(|A > B|, |B > A|)}{\sum (|A > B| + |A \equiv B| + |B > A|)} \quad (7.1)$$

The error rate quantifies the chance of reaching a wrong conclusion about a run pair. Note that due to its definition, the error rate can never be more than 50%.

However, a low error rate does not only indicate a high confidence in the conclusion that run A is better than run B, a measure can also have a low error rate simply because it rarely concludes that two runs are different. Hence, the number of times runs are deemed to be equivalent is also of interest as it reflects on the power of a test collection to discriminate among retrieval runs. The discriminative power is quantified by the *proportion of ties* (PT), which is defined as the number of times two runs are considered as equal ($|A \equiv B|$) divided by the total number of comparisons:

$$PT = \frac{\sum |A \equiv B|}{\sum (|A > B| + |A \equiv B| + |B > A|)} \quad (7.2)$$

The higher the proportion of ties, the lower the discriminative power of the performance comparison.

Reliability of Performance Comparisons

In order to quantify the stability of the *IAPR TC-12 Image Benchmark*, we took all 150 non-visual runs submitted to *ImageCLEFphoto 2006* and compared each run with every other, resulting in 11,175 comparisons. Further, we created 20 different topic sets (each consisting of 30 queries that were randomly selected from the *ImageCLEFphoto 2006* topics), yielding a total of 223.500 pair-wise comparisons for each performance measure used in that evaluation event. Figure 7.8 displays the average error rates of the four lead measures of the *IAPR TC-12 Image Benchmark* with respect to fuzziness values between 0 and 20%.

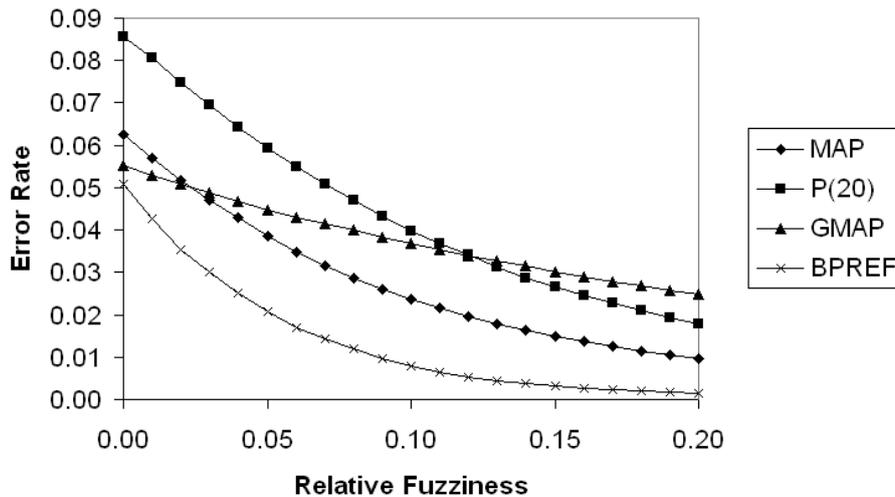


Figure 7.8: Error rates.

There is a consistent decrease in error rate as the fuzziness value increases, whereby *bpref* and MAP provide the highest levels of confidence in their results, while $P(20)$ shows the highest error rates. This also concurs with other studies such as [34, 372, 381, 466, 473, 510].

The error rates for all benchmark measures are relatively low, indicating a high reliability of the performance comparisons. Using MAP, for example, the decision that run A is at least 5% better than run B would only lead to an error in 3.9% of all comparisons, while the error associated with the decision that run A outperforms run B by at least 10% only amounts to 2.38% (see Table 7.18 for the error rates of all the measures at these fuzziness values).

ER for...	f = 0.05	f = 0.10
MAP	3.86 %	2.38 %
P(20)	5.92 %	3.97 %
GMAP	4.47 %	3.69 %
bpref	2.08 %	0.80 %

Table 7.18: Error rates associated with fuzziness values of 5% and 10%.

While larger fuzziness values decrease the error rate, they also decrease the discrimination power of the measures of the test collection.

Discriminative Power

The cost associated with increasing the difference is that fewer conclusions can be drawn since more methods are considered equal. Thus, larger fuzziness values do not only decrease the error rate, they also decrease the discrimination power of the measures. Figure 7.9 quantifies the effect of the fuzziness value on the discrimination power of the measures. Figure 7.9 quantifies the effect of the fuzziness value on the discrimination power of the measures.

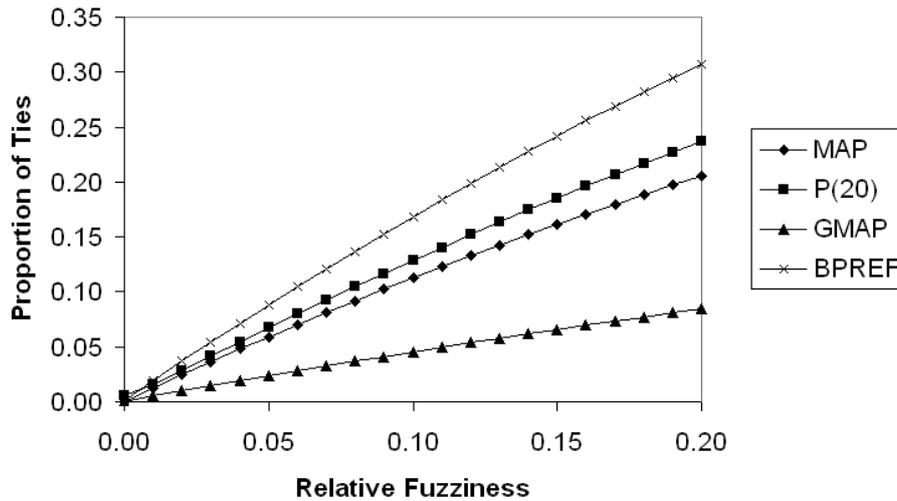


Figure 7.9: Proportion of ties.

Unsurprisingly, *bpref* as the measure providing the highest confidence in its experimental conclusions also shows the highest proportion of ties, while GMAP offers the highest discriminative power of all the measures. Overall, however, all measures again depict very low numbers of proportion of ties: only 5.86% of the comparisons using MAP would conclude that the difference between two runs is less than 5%, while only 11.3% would do so for a fuzziness value of 10% (see Table 7.19 for the proportion of ties of all the measures at these fuzziness values).

Difference	f = 0.05	f = 0.10
MAP	5.86 %	11.30 %
P(20)	6.78 %	12.92 %
GMAP	2.40 %	4.55 %
bpref	8.82 %	16.67 %

Table 7.19: Proportion of ties for fuzziness of 5% and 10%.

Test Collection Quality

How good is the *IAPR TC-12 Image Benchmark*? Ideally, a test collection would provide measures exhibiting a high reliability in their performance comparisons as well as strong discriminative power. Hence, for a good test collection, both the values for error rates (ER) and proportion of ties (PT) should be small [372].

In reality, however, a trade-off exists between these measures, and since fixed fuzziness values imply different trade-offs for different metrics, we vary $f = [0, 0.01, 0.02, \dots, 0.20]$ and plot PT and ER in order to evaluate the stability of the collection (see Figure 7.10).

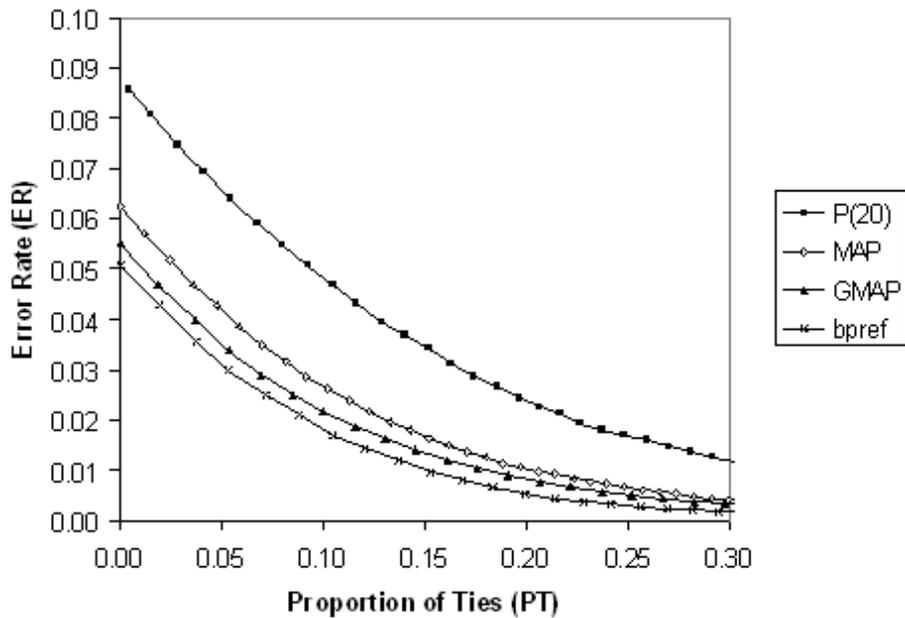


Figure 7.10: ER-PT curves.

The low values for both ER and PT are an indication for the high quality of the *IAPR TC-12 Image Benchmark* as a test collection in general, while an individual analysis of the measures shows that *bpref*, GMAP and MAP exhibit very high stability, with P(20) lagging slightly behind and being the least stable measure. These results are also in agreement with those reported in [372, 373].

Comparison with Other Collections

To allow for an objective evaluation of collections that were built using the TREC methodology, Voorhees [473] suggests that test collections should be compared by the minimum fuzziness value that is required to allow for $ER \leq 5\%$. Table 7.20 provides this information together with the respective proportion of ties for each of the performance indicators of the *IAPR TC-12 Image Benchmark*.

Measure	Difference (f)	Ties
MAP	2.33 %	2.87 %
P(20)	7.14 %	9.42 %
GMAP	2.35 %	1.19 %
bpref	0.10 %	0.30 %

Table 7.20: Required fuzziness values for $ER \leq 5\%$.

Again, the data indicate that a fuzziness value as small as $f = 2.33\%$ is sufficient to satisfy $ER \leq 5\%$ for MAP. The low fuzziness value required also yields a low proportion of ties (2.87%) and therefore allows for a high discrimination power of the performance indicator used. Now, how do these values compare to other collections? Table 7.21 provides a comparison with the fuzziness values reported for several TREC and NTCIR collections [372, 373, 473] to reach the 5% error rate limit when using MAP.

Collection	Required fuzziness (f)
TREC-3	4.1%
TREC-4	5.5%
TREC-5	6.1%
TREC-6	7.2%
TREC-7	4.4%
TREC-8	4.3%
TREC-9	5.4%
NTCIR-3 (Chinese)	11.0%
NTCIR-3 (Japanese)	11.0%
NTCIR-3 (English)	14.0%

Table 7.21: Minimum fuzziness values of other collections for $ER \leq 5\%$.

The larger differences required in other IR test collections to be confident in conclusions confirm the high stability of the *IAPR TC-12 Image Benchmark*. This

outstanding result may be credited to the diligent creation of the image database, the careful selection of query topics, and the fact that, unlike in many other test collections, complete relevance assessments were carried out.

7.4.2 Task Difficulty

One of the key contributions within this research was the development of a measure to quantify topic difficulty for concept-based image retrieval in order to assist with the topic development process and to create a balanced topic set that is neither too difficult nor too easy for existing techniques, while it is also considered as crucial to increase the yearly difficulty levels to keep up the challenge for returning participants (see Section 5.1).

To evaluate whether these challenges have been achieved using the novel difficulty measure, we compare the difficulty levels and results of *ImageCLEFphoto 2006* with the ones from the previous *ImageCLEF* ad-hoc retrieval tasks.

Year	avg d	avg MAP	avg P(20)
2004	2.09 (0.37)	0.3720 (0.1701)	0.4001 (0.2382)
2005	2.36 (0.32)	0.3171 (0.1482)	0.3750 (0.1563)
2006	2.92 (0.69)	0.1584 (0.1157)	0.2280 (0.1392)

Table 7.22: Topic difficulty in ImageCLEF 2004-2006.

Table 7.22 shows the development of the average topic difficulty levels (with their standard deviations in parenthesis) and the average precision values achieved by participants of *ImageCLEF* from 2004 to 2006. Topics have, indeed, consistently become more difficult each year, however MAP values have also dropped at a similar rate as the difficulty has increased. This could be due to a number of reasons, including the use of similar IR techniques each year, a more visually challenging collection, and there being incomplete representations provided with the collection.

Therefore for *ImageCLEFphoto 2007*, we are planning to create topics with an average difficulty level of around $d = 3.0$ again to investigate whether methods for VIR from generic photographic collections have advanced within the last 12 months or not (see also Section 7.4.5).

7.4.3 Performance Measures

Next, we used Kendall’s rank correlation coefficient (τ) to compare the system rankings between the measures used for evaluation. Correlations of $\tau < 0.8$ generally reflect noticeable changes in the rankings and suggest that the measures have a different evaluation emphasis, whereas correlations of $\tau > 0.9$ can be considered as equivalent [467]. Hence, our set of performance measures would ideally exhibit correlations of $\tau < 0.9$, because if any two measures had shown a correlation of $\tau > 0.9$, one of them would have been redundant and could have been dropped as they both would have expressed the same evaluation emphasis.

Kendall (τ)	MAP	P(20)	BPREF	GMAP
MAP	N/A	0.852	0.797	0.741
P(20)	0.852	N/A	0.735	0.742
BPREF	0.797	0.735	N/A	0.854
GMAP	0.741	0.742	0.854	N/A

Table 7.23: Correlation of performance measures.

Table 7.23 shows that significant correlations between 0.74 and 0.85 exist at $p \leq 0.01$ between all the measures above. As a consequence, the fact that all our measures show correlations of $\tau < 0.9$ (and most of them even $\tau < 0.8$) indicates that:

- the choice for a particular measure used at *ImageCLEFphoto 2006* did, in fact, affect system ranking;
- the set of measures chosen for *ImageCLEFphoto 2006* allowed for a non-redundant evaluation of retrieval performance, with each of the measures emphasising different aspects of retrieval and therefore complementing each other.

We therefore proposed to reuse the same set of measures for *ImageCLEFphoto 2007*, which was approved by the majority of the participants (see Section 7.4.4).

7.4.4 Feedback From Participants

A vital component for the success of any evaluation event is the feedback of its participants. Evaluation events are often compared by the number of participants they attract, and only if the event is well organised and offers interesting tasks, would researchers return to participate the following year and new participants be attracted. Hence, we created a feedback form (which was subsequently distributed to all participating and non-participating groups) prior to CLEF and asked for comments regarding the organisation of the evaluation event in general and the specific benchmark components in particular.

Document Collection

All participants unanimously agreed that the *IAPR TC-12 Image Benchmark* provided an appropriate test collection for *ImageCLEFphoto 2006*, representing a realistic set of real-life still natural images and being easy to access and download (only one participant mentioned that the copyright status was not very clear).

Moreover, the majority of participants also judged the quality of the logical image representations in the collection between good and excellent, with some participants taking a rather neutral position. Most participants also approved the idea of a parametric benchmark architecture in general and the fact that only a subset of the logical image representations had been provided in order to make the task even more realistic, and all participants would like to experiment with the *IAPR TC-12 Benchmark* at *ImageCLEFphoto 2007* again.

Query Topics

Most of the participating groups considered the number and difficulty of the topics as appropriate and agreed with the topic creation process being based on several query dimensions. Only two participants pointed out that they found the topics a bit too contrived, while two other participants would have liked to see more than 60 topics for evaluation.

However, the topics for the additional visual subtask were not perceived as very useful by several groups, which is also indicated by the low number of submissions (only two out of 36 registered groups eventually submitted). Some groups mentioned in their feedback that they could not submit due to lack of time; the generally low results for this task might have also discouraged several groups from submitting their results. On the other hand, there were twice as many groups that submitted purely content-based runs to the general *ImageCLEFphoto* task, which raises the question whether this additional visual sub-task had been sufficiently promoted before the event. However, most participants agreed that purely visual topics should be part of the standard topics for *ImageCLEFphoto*, rather than a stand-alone task.

Relevance Assessments and Performance Measures

As far as relevance assessments are concerned, using the pooling method combined with ISJ to complement the ground-truth with further relevant images was considered as appropriate by all participants (although one group was concerned with the amount of work involved for the organisers).

The proposed set of measures *MAP*, *P(20)*, *GMAP*, and *bpref* was also accepted by the majority of the groups, with all the participants agreeing on keeping *MAP* as the leading measure to express retrieval performance. However, two participants expressed their concern about *P(20)*, and one participant questioned whether *GMAP* and *bpref* should be considered for the future, yet both without explaining any specific reasons for their disliking of these measures nor providing any alternative solutions.

Event Organisation

Finally, all participants unanimously agreed that *ImageCLEFphoto 2006* was very well organised, that they received a satisfactory level of communication from the organisers and that the web site had all the information required for the task. Only one participant pointed out that the submission instructions were not clear, while one other participant did not find the submission system suitable.

Overall, participants agreed that they found *ImageCLEFphoto 2006* very useful and all groups (except for one still undecided group) indicated that they would participate at *ImageCLEFphoto 2007* again.

Non-Participating Groups

The majority of the groups that had registered (but eventually failed to officially submit to *ImageCLEFphoto 2006*) mentioned that they had not been able to complete their retrieval experiments on time and/or that their results had not been good enough to be presented. Some also stated that they had only registered in order to be granted access to the test collection for the time being, but were thinking of participating in future events such as *ImageCLEFphoto 2007*.

7.4.5 Future Prospects

Information retrieval benchmarks are generally considered to be an ongoing and incremental process, and thus the documentation of *ImageCLEFphoto 2006* would not be complete without reporting on its influence of its succeeding event *ImageCLEFphoto 2007*. Based on the experience and the feedback retrieved in 2006 and on discussions with participants after CLEF, future prospects for 2007 will include the following.

Document Collection

The *IAPR TC-12 Benchmark* will again form the basis for the VIR experiments, whereby only realistic parts of the logical image representations will be released in 2007: title, notes, location, and date fields (*i.e.* the descriptions that typical users might add to their own photographs). In addition to the English and German representations, we are planning to also generate a set of Spanish image representations as well as one subset using a randomly selected language for each image. Evaluation of ad-hoc retrieval from lightly annotated images is expected to address several novel research questions including:

- do traditional retrieval methods still work with short image representations?

- how significant is the choice of the retrieval language?
- how does retrieval performance compare to retrieval from fully annotated images in 2006?

Since the involvement of visual retrieval techniques will become more important, we aim to attract more visually oriented methods in addition to the currently predominant concept-based approaches to further approach and narrow the semantic gap from both sides, TBIR and CBIR.

Query Topics

According to the participants' feedback from 2006, the query topics in 2007 will:

- again be based on the updated *viventura* log file (to create realistic topics);
- reuse some of the topics from 2006 (for a comparison with retrieval using the description field, and to investigate how much improvement can be gained one-year on);
- be controlled by the topic difficulty measure;
- be created against a number of dimensions such as the estimated size of the target set, geographic constraints or the level of how “visual” they appear.

Participants will only receive the topic titles and three sample images, but no narrative descriptions to avoid confusion. The sample CBIR systems FIRE and GIFT will also be available again, and we might even provide the output of visual baseline runs. Translations only for topic languages that were also used in 2006 will be provided for 2007 as well. These are: English, German, Spanish, Italian, French, Portuguese, Russian, Polish, Japanese, Simplified and Traditional Chinese. Visual topics will thereby be part of the standard ad-hoc set. Should participants wish to investigate any other language, they will have to provide their own translation.

Further novel ideas include that participants could choose their own sample images for QBE and that participants could be asked to submit a number of topic candidates themselves.

Relevance Assessments and Performance Measures

Both relevance assessments and performance measures will remain unchanged for 2007: the use of the pooling method combined with ISJ and the same set of measures (*MAP*, *P(20)*, *GMAP*, and *bpref*). New ideas include the ranking of systems by the average rank of these measures, and to further involve the participating groups in the relevance assessment process.

7.5 Summary

This chapter reported on *ImageCLEFphoto 2006*, the first evaluation effort for (multilingual) VIR from a generic photographic collection (*e.g.* photographs of holidays and events).

First, after a general introduction to ad-hoc retrieval tasks at *ImageCLEF*, the motivation for providing *ImageCLEF* with the resources and functionality of the *IAPR TC-12 Image Benchmark* was presented, followed by a chronological description of the organisation and realisation of the evaluation event from January to December 2006. In particular, it was highlighted how the individual benchmark components were generated and used in the light of *ImageCLEFphoto*: this included the image collection and the query topics as well as the relevance judgments and the choice for a particular set of performance measures.

ImageCLEFphoto 2006 saw the submission of 157 system runs by 12 participating groups from 10 different countries. A description for each of the systems used in the evaluation was provided, together with an analysis of their retrieval performance with respect to several submission parameters and topic dimensions. Some of the findings include:

- a combination of visual and textual features generally improves retrieval effectiveness;
- visual features often work well for more visual queries;
- multilingual image retrieval is as effective as monolingual retrieval;

- feedback and query expansion can help to improve retrieval effectiveness.

Although some of these trends had been shown for other domains before, *ImageCLEFphoto 2006* was the first large-scale evaluation event to actually investigate these also for the domain of multilingual retrieval from a generic photographic collection. Finally, an analysis of the event was provided too, including the evaluation of the task difficulty, the choice of performance measures, the feedback of participating groups and, based on it, the future prospects for *ImageCLEFphoto 2007* and onwards.

After the image retrieval community had been calling for resources similar to those used by TREC in the document retrieval domain, *ImageCLEF* has begun to provide such resources also within the context of VIR in order to facilitate standardised laboratory-style testing of (predominately concept-based) image retrieval systems. By running evaluation tasks which are modelled on scenarios found in multimedia use today, the barriers between research interests and real-world needs have been addressed.

These resources now also include a benchmark suite for retrieval from generic photographic collections, a domain that had lacked such resources for retrieval evaluation for a long time, although it had been estimated to be likely to become of increasing interest to researchers with the growth of the desktop search market (and the popularity of tools such as *Flickr*). By joining the *IAPR TC-12 Image Benchmark* with the *ImageCLEFphoto* ad-hoc retrieval task, the need for evaluation events in this domain has now been satisfied, and the gap has finally been filled.

Chapter 8

Conclusion

This last chapter summarises the original work presented in this dissertation, recalls the scientific contributions and explains the limitations of this research.

8.1 Summary

This dissertation has investigated the system-centred evaluation of (multilingual) VIR from generic photographic collections and is composed of eight chapters, including the introduction and this concluding chapter; the remaining six main chapters are briefly summarised below.

Characteristics and Processes of Visual Information Retrieval

In recent years, there has been an increasing amount of literature on the main concepts and challenges of VIR. An unsolved problem to date is the so called semantic gap, which is the discrepancy between the information that one can automatically extract from visual data and the interpretation of the same data for a user in a given situation.

Research endeavours to bridge the semantic gap have thereby taken two contrary approaches: content-based image retrieval (CBIR) is based on purely visual features (such as colour, texture and shape) that can be directly extracted from images, while concept-based image retrieval (TBIR) relies on meta-data or additional alphanumeric representations associated with the images to express their semantics.

We provide an analysis and classification of visual information queries, similarity measures and the result generation process.

Analysis and Evaluation of Visual Information Retrieval

For the field of visual information search to advance, objective evaluation to identify, compare and validate the strengths and merits of different systems is essential. Uniform sets of data, queries, relevance judgments and measures of performance are therefore needed to provide a standardised platform (called benchmarks or test collections) to carry out such an evaluation, together with evaluation events to also attract researchers to make use of these components.

Such benchmarks have recently been developed (and evaluation events have been organised) for several domains of VIR, including the retrieval from historic or medical collections, object recognition and automatic annotation tasks for general collections as well as for specific ones like coin images or radiographs, user-centred evaluation of systems and also in related fields such as video retrieval, cross-language information retrieval and multimedia retrieval from structured (XML) collections. No efforts, however, had considered the evaluation of multilingual retrieval from generic photographic collections (*i.e.* containing everyday real-world photographs akin to those that can frequently be found in private photographic collections as well, *e.g.* pictures of holidays and events).

The goal of this research was therefore fill this gap by designing and implementing the required resources to carry out such an evaluation: the *IAPR TC-12 Benchmark*. These resources include (1) the design and development of a standardised image collection for this domain, (2) the creation of representative search topics and relevance judgments to associate a ground-truth of relevant images for each of these topics, (3) a set of performance measures to quantify, rank and evaluate the results, and (4) the organisation of an evaluation event to practically apply these components and provide them to the research community.

Data Design and Engineering

A core component of image retrieval benchmarks is a set of images that are representative of a particular domain. Finding such resources for general use is often difficult, not least because of copyright issues which restrict the distribution and future accessibility of data. This is especially true for visual resources that are often expensive to obtain and subject to limited availability and access for the research community.

We therefore report on the creation of an image collection called the *IAPR TC-12 image collection*, which we specifically designed and implemented to deal with the lack of resources for evaluation of VIR from generic photographic collections. The goal was to provide:

- a collection of general, real-world photographs suitable for a wider range of evaluation purposes;
- images with associated written information representing typical textual metadata to allow for the exploration of the semantic gap;
- semantic image descriptions in multiple languages as such real-life collections are inherently multilingual;
- a data set that is free of charge and copyright restrictions and therefore available to the general research community.

To achieve these goals, we first specified the requirements for the creation of such a collection, including the definition of rules for the *image selection* and *annotation* processes, which would subsequently allow for the strict control over the consistency and quality within all aspects of collection creation. We then acquired access to an image database of general photographs (photos of travel destinations, tourists and events) and, following the rules, we selected 20,000 images and annotated them in three languages: English, German and Spanish.

Task Creation and Visual Information Complexity

The second key component of the *IAPR TC-12 Image Benchmark* is a set of representative *search requests* (query topics). The specific goal was to develop a natural, balanced topic set accurately reflecting real world user statements of information needs for retrieval from the *IAPR TC-12 image collection*.

In general, such statements of user information needs are created against certain task parameters (dimensions) to allow for some control over the topic creation process. Thus, we first identified the dimensions specific to retrieval evaluation using the *IAPR TC-12 image collection*, which include the total number of topics provided and for each topic: the estimated number of relevant documents (images), the topic scope (*e.g.* broad or narrow, general or specific) and origin, the use of geographic constraints, the representation completeness of relevant images, the estimated difficulty, the likelihood of retrieval success using visual features only, and supplementary task creation parameters such as additional text retrieval challenges and feedback from participants.

To base the topic creation process on realistic user information needs, we first implemented a logging function for a web-based interface to the *IAPR TC-12 image collection* and subsequently analysed the search behaviour and query patterns specific to retrieval from this database. Based on the topic candidates following the results from the log file analysis, we then created a set of representative query topics against the aforementioned query dimensions.

No work had considered the topic difficulty for TBIR. To be able to also balance the query topics for difficulty, we designed a novel measure to quantify topic difficulty for TBIR based on both linguistic features of the topic and statistical information gained from the corresponding document collection. Experimental validation and a comparison with other approaches showed that the novel measure displays a strong negative correlation between topic difficulty and system effectiveness and gives an upper boundary of the correlation which can be achieved using a costly manual approach. We purport that having such an accurate measure en-

ables the creators of TBIR evaluation events to carefully select topics, making topic difficulty one of the most significant dimensions in the topic creation process.

Parametric Benchmark Design and Architecture

To facilitate the incremental development as well as the ongoing maintenance and administration of the benchmark collection (*i.e.* images and their corresponding semantic descriptions) and the creation and administration of the representative query topics, we designed and implemented a benchmark administration system.

The most significant benefit of this novel benchmark architecture can be found in its *parametric* nature, which allows for a fast adaptation to changed retrieval requirements or new evaluation needs. Collection parameters include the size of the collection, the contents and complexity of images and their geographic or temporal distribution. Examples for image representation parameters are their type, format, language, completeness and the quality level of orthography. The benchmark administration system thereby also supports this parametric benchmark paradigm and facilitates the quick reaction to such changes in research direction by simply altering the parameters and the subsequent regeneration of the required subsets.

Further merits of the benchmark administration system include the facilitation of the incremental collection development, the guidance of the creation, administration, translation and generation of representative search topics, and the the efficient execution of relevance assessments.

System Evaluation and Analysis

The benchmark components summarised in the preceding sections certainly provide excellent resources to the information retrieval and computational vision communities to facilitate standardised laboratory-style testing of (predominately concept-based) image retrieval systems. However, such resources can only prove beneficial to research if they are actually used in evaluation events as well.

Hence, we have used the *IAPR TC-12 Image Benchmark* in a multilingual ad-hoc image retrieval task (called *ImageCLEFphoto 2006*) at the *ImageCLEF 2006*

evaluation campaign. Reasons for the choice of a multilingual environment as evaluation platform include:

- the task scenario offered by its ad-hoc retrieval task, which is very similar to that modelled by the *IAPR TC-12 Benchmark*;
- the broad range of audience and participation in prior *ImageCLEF* campaigns;
- the multilingual evaluation environment provided by *ImageCLEF*, which represents the most realistic model for evaluation of retrieval from general photographs since such real-life collections are inherently multilingual;
- the lack of the resources to organise an evaluation event on our own.

ImageCLEFphoto 2006 was the first evaluation event for (multilingual) ad-hoc retrieval from generic photographic collections, and we organised it following an adapted methodology that the Text REtrieval Conference (TREC) had successfully used in the text retrieval domain. The annual cycle of events thereby comprises (in chronological order): the call for participation, registration, document release, topic release, result submission, the creation of relevance assessments, result generation, the actual evaluation event, and the final publication of methods and results.

We highlight how the individual benchmark components were generated and used in the light of *ImageCLEFphoto*, including the image collection and the query topics as well as the relevance judgments and the choice for a particular set of performance measures. We analysed more than 150 system runs submitted by 12 participating groups from 10 different countries. Some of the findings include that:

- a combination of visual and textual features generally improves retrieval effectiveness;
- visual features often work well for more visual queries;
- multilingual image retrieval is as effective as monolingual retrieval;
- feedback and query expansion can help to improve retrieval effectiveness.

ImageCLEFphoto 2006 was the first large-scale evaluation event ever to actually investigate these findings for the domain of multilingual retrieval from a generic photographic collection.

We further analysed the test collection and the evaluation event itself, and, based on our results and feedback from participants, we claim that:

- the benchmark provides performance comparison of retrieval runs with high reliability and discrimination power (as quantified by the error rate and the proportion of ties);
- the difficulty of the retrieval tasks was appropriate (as quantified by the topic difficulty measure);
- the selection of performance measures was useful (as indicated by their correlation values);
- our methodology of parametric benchmarking for image collection and topic creation was validated and approved by the research community;
- we successfully addressed the barriers between research interests and real-world needs by organising an evaluation task modelled on a scenario found in multimedia use today.

Last but not least, and based on all the above, we purport that we successfully repaired the lack of evaluation for (multilingual) visual information retrieval from generic photographic collections.

8.2 Main Achievements

This section recalls the main scientific contributions of the research presented in this dissertation. These contributions have already been indicated in Section 1.3.3 and have been detailed in several chapters afterwards. The chapters on data design and engineering (4), task creation and visual information complexity (5), parametric

benchmark design and architecture (6) and on system evaluation and analysis (7), in particular, bear the content of these scientific contributions (see Figure 8.1).

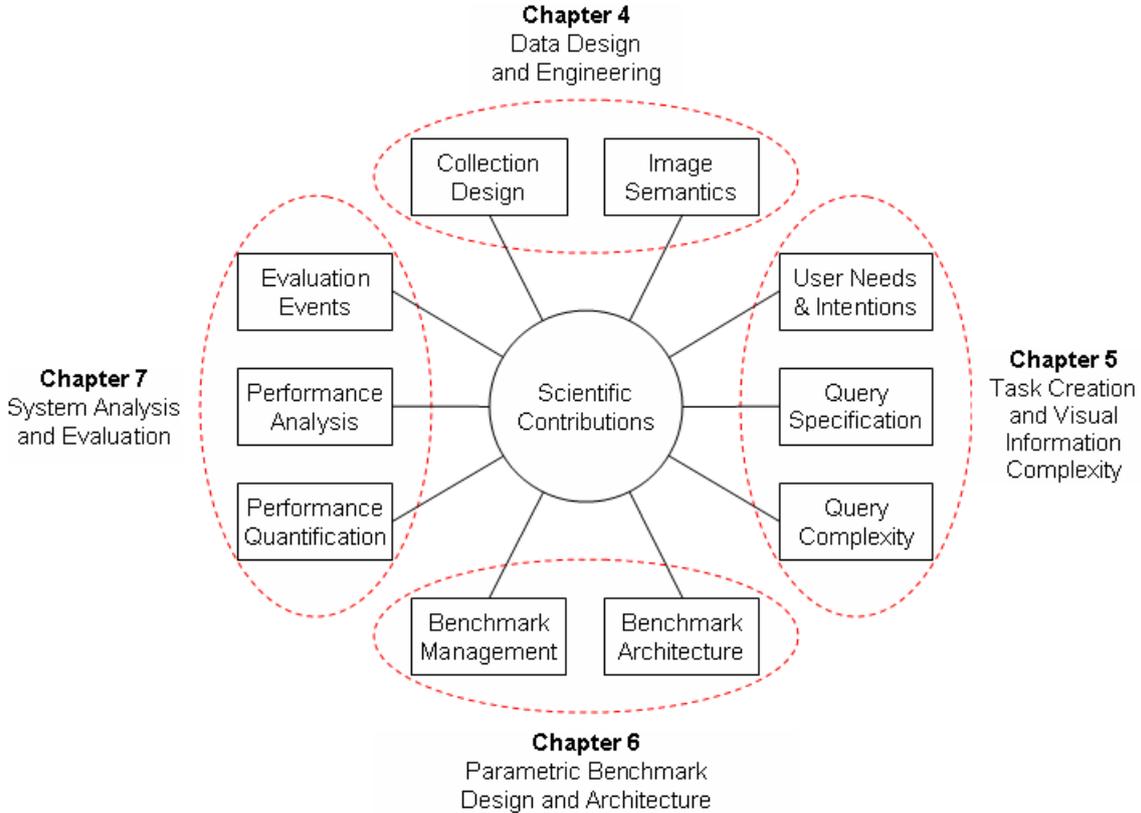


Figure 8.1: Scientific contributions.

We have studied and made contributions to the *design* of parametric test *collections*, the universality of *image semantics* and logical image representations across different languages and world views, the matching of *user intentions* and *query specifications*, *query complexity*, *benchmark management* and *architecture*, *performance quantification* and *analysis*, and the design of *evaluation events*. These contributions make possible a systematic calibration and comparison of system performance for (multilingual) VIR from generic photographic collections.

We have further shown that, with VIR, it is not just a matter of issuing queries against a database and obtaining results, but rather it requires the analysis of a multitude of variables and factors. The work presented in this dissertation therefore also enables a deeper understanding of the complex conditions and constraints

associated with visual information identification, the accurate capturing of user requirements, the correct expression of user queries, the complexity of queries, the execution of searches, and the reliability of performance indicators.

8.3 Limitations and Future Research

Although the topic creation process had been based on topic candidates derived from a log file analysis, and topics had been created against a number of dimensions to allow for additional control, there are still always negative voices that claim that topics were too contrived and not realistic at all. We therefore also recommend that further research be undertaken in the area of topic development and result generation.

More information on what types of searches users typically perform in the domains would, in general, help to establish a greater degree of accuracy in creating realistic topics for evaluation events. In the case of the *IAPR TC-12 Benchmark*, such investigation could be accomplished by re-analysing the log files from online access to the collection. While the original analysis was only based on 980 unique queries, the file has now accumulated more than 5,000 entries¹, representing a much more significant sample for investigation.

One drawback of the methodology for topic creation and management can be seen in the huge amount of work involved for the organisers of an evaluation event. Not only does the identification of topic candidates and the development of representative topics against several dimensions take up a considerable amount of time, but the translation of topics, the selection of sample images for query-by-visual-example approaches, and especially the carrying out of relevance assessments can also be very time-consuming and cumbersome tasks.

Solutions to ease the amount of work for organisers include (1) the idea to let participants choose their own sample images to start their visual queries or (2) to make it a requirement for participating groups at evaluation events to provide a

¹As of 27 April 2006.

number of topic candidates themselves (as practised at INEX Multimedia) and/or to also assist with relevance assessments. The question arises whether this would have any negative effects on the number of participants (*e.g.* INEX Multimedia could not attract more than five participants thus far).

It has further been suggested to save time and effort by replacing the proposed method for the difficulty estimation of topics, and using alternative automatic approaches instead. However, this would come at a cost of lowering correlation and ultimately being less successful at predicting system effectiveness, a compromise too severe to accept as we consider the quantification of topic difficulty as one of the key dimensions within the topic creation process.

Glossary

This chapter contains a list of all the abbreviations used in this dissertation.

ACM Advanced Computing Machinery

AGFA Actien-Gesellschaft für Anilin-Fabrikation (an imaging company)

ALOI Amsterdam Library of Object Images

AMI Augmented Multi-party Interaction (video retrieval evaluation event)

AP Average Precision

API Application Programming Interface

ARGOS Evaluation Campaign for Surveillance Tools of Video Content (French)

ART Angular Radial Transform

AUC Area Under Curve

AVG Average

BG Background

BPREF Binary Preference

CBIR Content-Based Image Retrieval

CBIRS Content-Based Image Retrieval System

CCV Colour Coherence Vector

CCTV4 China Central Television 4

CEA Commissariat à l'Énergie Atomique (French Atomic Energy Commission)

CFW Collection Frequency Weight

CGI Common Gateway Interface

CHI Computer–Human Interaction

CIE Commission Internationale de l'Eclairage (International Commission of Illumination)

CIS Coin Images Seibersdorf

CLEF Cross–Language Evaluation Forum

CLIR Cross–Language Information Retrieval

CMY Cyan, Magenta, Yellow (color space)

CNN Cable News Network

COIL Columbia University Object Image Library

CSS Curvature Scale Space

CT Computer Tomography

CV Computer Vision

DARPA Defense Advanced Research Project Agency

DB Database

DBMS Database Management System

DCU Dublin City University

DDL Description Definition Language

DFR Divergence From Randomness

DS Description Scheme

EER Equal Error Rate

EM Expectation Maximisation

ETISEO Evaluation du Traitement et de l'Interprétation de Séquences Vidéo
(Video Understanding Evaluation)

FD Fourier Descriptor

FG Foreground

FIRE Flexible Image Retrieval Engine

GIFT GNU Image Finding Tool

GIR Geographical Information Retrieval

GIS Geographical Information System

GMAP (geometric) Mean Average Precision

GNU GNU is Not Unix

GPL General Public License

GRF Gibbs Random Field

GUI Graphical User Interface

HEAL Health Education Assets Library

HMMD Hue Maximum Minimum Difference (colour space)

HSB Hue, Saturation, Brightness (colour space)

HSV Hue, Saturation, Value (colour space)

HTML Hypertext Markup Language

HTTP Hypertext Transfer Protocol

IAPR International Association for Pattern Recognition

IBF International Boxing Federation

IBM International Business Machines

ID Identification, unique identifier

IDF Inverse Document Frequency

IEEE Institute of Electrical and Electronics Engineers

INEX Initiative for the Evaluation of XML Retrieval

IR Information Retrieval

IRMA Image Retrieval in Medical Applications

ISJ Interactive Search and Judge (pooling method)

JPEG Joint Photographic Experts Group

KLT Karhunen-Loeve Transform

LBC Lebanese Broadcasting Corporation

LCD Liquid Crystal Display

LNCS Lecture Notes in Computer Science

LSE Least Significant Element

LSI Latent Semantic Indexing

LSW Least Significant Word

LTU Look That Up Technologies (company)

MAP (arithmetic) Mean Average Precision

MARS Multimedia Archival and Retrieval System(s)

MFD Minkowski-Form Distance

MGRF Markov-Gibbs Random Field

MIR Mallinckrodt Institute of Radiology

MIRA Multimedia Information Retrieval Applications

MIRC Medical Image Resource Centre

MIT Massachusetts Institute of Technology

MPEG Moving Picture Experts Group

MRF Markov Random Field

MRI Magnetic Resonance Imaging

MRSAR Multi-Resolution Simultaneous Auto Regressive texture model

MSE Most Significant Element

MRR Mean Reciprocal Rank

MSN Microsoft Network

MSNBC Microsoft Network / National Broadcasting Company

MSW Most Significant Word

MT Machine Translation

MTF Move To Front (pooling method)

NBC National Broadcasting Company

NEC Nippon Electric Company

NII National Institute of Informatics

NIST National Institute of Standards and Technology

NLP Natural Language Processing

NTCIR NII Test Collection for IR Systems

NTDTV New Tang Dynasty Television

NTU National Taiwan University

OWL Web Ontology Language

PA Place Adjunct

PCA Principal Component Analysis

PEIR Pathology Educational Instructional Resource

PETS Performance Evaluation of Tracking and Surveillance

PHP PHP: Hypertext Preprocessor

PNG Portable Network Graphics

PR Pattern Recognition

PR graph Precision *vs.* Recall graph

QBE Query by Example

QBIC Query by Image Content

QBK Query by Keyword

QBS Query by Sketch

QFD Quadratic Form Distance

RDF Resource Description Framework

RF Relevance Feedback

RGB Red, Green, Blue (colour space)

RIA Reliable Information Access (workshop)

RISAR Rotation-Invariant Simultaneous Auto-Regressive texture model

ROC Receiver Operator Characteristic

SAC St. Andrews Collection of historic photographs

SAR Simultaneous Auto-Regressive texture model

SGML Standard Generalised Markup Language

SIGIR Special Interest Group on Information Retrieval

SIR Semantic Image Retrieval

SOIL Surrey Object Image Library

SOM Self Organising Map

SPEC Standard Performance Evaluation Corporation

SQL Structured Query Language

TA Time Adjunct

TBIR Text-Based Image Retrieval

TBIRS Text-Based Image Retrieval System

TC Technical Committee

TF Term Frequency

TFM Ternary Fact Model

TPC Transaction Processing Performance Council

TREC Text REtrieval Conference

TUC Technical University Chemnitz

URL Uniform Resource Locator

USA United States of America

USD United States Dollar

VIPER Visual Information Processing for Enhanced Retrieval

VIR Visual Information Retrieval

VIRS Visual Information Retrieval System

VOC Visual Object Classes

WBA World Boxing Association

WBC World Boxing Council

WBO World Boxing Organisation

WJM Wolf Jolion Metric

WWW World Wide Web

XML eXtensible Markup Language

Notation

This chapter explains the mathematical notation used throughout this dissertation to keep it as homogeneous as possible. It also provides a link to where the symbol was used for the first time to get further information on it.

$|\cdot|$ Cardinality of a set (introduced in Equation 5.4)

A Similarity matrix (introduced in Equation 2.8)

AP Average precision (defined in Equation 3.3)

AP_i Average precision for topic i (introduced in Equation 3.4)

a_{ij} Similarity between i and j (introduced in Equation 2.8)

B The number of bins in a histogram (introduced on Page 70)

b Tuning constant (used in Equation 2.5)

$bpref$ Binary preference (defined in Equation 3.11)

C Covariance matrix of feature vectors (introduced in Equation 2.9)

$cov(X, Y)$ Covariance of two variables X and Y (used in Equation 5.17)

\mathcal{D} Set of documents in a collection (introduced on Page 67)

D Number of documents in a collection (introduced in Equation 2.2)

D_j Text document in a collection (introduced on Page 67)

$D(I, J)$ Distance between images I and J (introduced on Page 70)

d Topic difficulty (defined in Equation 5.16)

d_j Topic difficulty for iteration j (defined in Equation 5.15)

$d(\mathcal{T}, \mathcal{I})$ Topic difficulty of topic \mathcal{T} for image collection \mathcal{I} (defined in Equation 5.16)

df Document frequency (defined in Equation 5.14)

$dl(j)$ Document length, length of document D_j (introduced on Page 67)

$E(X)$ Expected value of variable X (used in Equation 5.17)

ER Retrieval experiment error rate (defined in Equation 7.1)

er Error rate (defined in Equation 3.13)

F_I Feature vector for image I (introduced in Equation 2.8)

f Fuzziness value (introduced on Page 310)

$f_i(I)$ Number of pixels in bin i of image I (introduced on Page 70)

$GMAP$ Geometric mean average precision (defined in Equation 3.5)

H Entropy of objects in an image (defined in Equation 4.2)

H_{max} Maximum entropy of objects in an image (defined in Equation 4.4)

$H(I, J)$ Hausdorff distance between images I and J (defined in Equation 2.10)

$\vec{h}(I, J)$ Directed Hausdorff distance from image I to J (defined in Equation 2.11)

\mathcal{I} Set of images in the collection (introduced in Equation 5.16)

I, J Images (introduced on Page 70)

i, j, k, l Counters for various purposes

$idf(i)$ Inverse document frequency of term t_i (defined in Equation 2.2)

K Number of topic elements (introduced in Equation 5.2)

K_1 Tuning constant (used in Equation 2.5)

L_1 Manhattan distance (described on Page 71)

L_2 Euclidean distance (described on Page 71)

$m_{p,q}$ Algebraic moment of order $p + q$ (defined in Equation 2.1)

MAP Un-interpolated mean average precision (defined in Equation 3.4)

MRR Mean reciprocal rank of relevant images (defined in Equation 3.10)

\mathcal{N} Set of images retrieved (introduced in Equation 5.5)

\mathcal{N}_D Set of retrieved images through direct hits (introduced in Equation 5.4)

N Number of images in the collection (introduced on Page 141)

$N_{\mathcal{O}}$ Number of objects in an image (described on Page 174)

$N_{\mathcal{O}}(i)$ Number of objects of type i in an image (introduced in Equation 4.3)

N_R Number of binary relations in an image (defined in Equation 4.5)

N_{Rmax} Maximum number of binary relations (defined in Equation 4.6)

N_T Number of object types in an image (introduced in Equation 4.2)

n Number of retrieved images (introduced in Equation 3.1)

n_{cc} Number of correctly classified images (introduced in Equation 3.14)

n_{co} Number of image types correctly classified at least once (introduced in Equation 3.14)

n_{cu} Number of images classified as unknown (introduced in Equation 3.14)

n_{cw} Number of wrongly classified images (introduced in Equation 3.14)

n_r Number of relevant images retrieved (introduced in Equation 3.1)

n_{rjn} The first r judged non-relevant images (introduced in Equation 3.11)

$ndl(j)$ Normalised length of document D_j (defined in Equation 2.3)

\mathcal{O} Set of objects in an image (introduced on Page 174)

P Precision (defined in Equation 3.1)

PT Proportion of ties (defined in Equation 7.2)

P_i Precision for image i (introduced in Equation 3.4)

$P(n)$ Precision after n images are retrieved (introduced on Page 141)

$P(r)$ R-precision (introduced on Page 142)

p, q Order (introduced in Equation 2.1)

p_i Likelihood of the occurrence of object i in an image (defined in Equation 4.3)

Q Total number of query topics (introduced in Equation 3.4)

\mathcal{R} Set of relevant images for all topic sentence elements (defined in Equation 5.3)

\mathcal{R}_j Set of relevant images for all topic sentence elements in iteration j (introduced in Equation 5.3)

\mathcal{R}_j^* Set of relevant images for the most significant topic element in iteration j (defined in Equation 5.13)

$\mathcal{R}_{j,k}$ Set of relevant images for the k^{th} topic element in iteration j (described on Page 216)

\mathcal{R}_R Set of relevant images retrieved (introduced in Equation 3.8)

R Recall (defined in Equation 3.2)

$R(n)$ Recall after n images are retrieved (described on Page 141)

r Number of relevant documents/images (introduced in Equation 2.2)

- r_k The k^{th} relevant document/image (introduced in Equation 3.11)
- $Rank_1$ Rank of the first relevant image (introduced on Page 147)
- $Rank_k$ Rank of the k^{th} relevant image (introduced on Page 148)
- $Rank_{rec}$ Reciprocal Rank of the first relevant image (defined in Equation 3.8)
- \overline{Rank} Average rank of relevant images (defined in Equation 3.9)
- S MUSCLE CIS Score (defined in Equation 3.14)
- \mathcal{T} Topic sentence (defined in Equation 5.2)
- \mathcal{T}_j Set of topic elements for the j^{th} iteration (mentioned on Page 217)
- T The transpose of a matrix (introduced in Equation 2.8)
- t Topic sentence element (defined in Equation 5.1)
- t_i Term i in a text document (introduced in Equation 2.2)
- t_k The k^{th} topic element (introduced in Equation 5.2)
- $tf(i, j)$ Term frequency, number of occurrences of term t_i in document D_j (described on Page 67)
- X, Y Variables used in Pearson's product moment correlation (introduced in Equation 5.17)
- x, y Coordinates (used in Equation 2.1)
- α Factor for vocabulary mismatch and incomplete and incorrect annotation (defined in Equation 5.4)
- β Factor for element ambiguity (defined in Equation 5.5)
- γ Annotation gap factor (defined in Equation 5.6)
- η Tuning constant (introduced in Equation 5.6)

θ Tuning constant (introduced in Equation 5.6)

λ Constant for calculation of *GMAP* (introduced in Equation 3.7)

μ_X Mean value of random variable X (introduced in Equation 5.17)

ρ Object repetitiveness (defined in Equation 4.1)

$\rho(X, Y)$ Pearson's product moment correlation (defined in Equation 5.17)

σ_X Standard deviation of variable X (used in Equation 5.17)

τ Kendall's rank correlation coefficient (described on Page 317)

Bibliography

- [1] *Proceedings of the 9th ACM International Conference on Multimedia (ACM Multimedia 2001)*, Ottawa, Canada, September 30 - October 5 2001. ACM Press.
- [2] M. Addis, K. Martinez, P. Lewis, J. Stevenson, and F. Giorgini. New Ways to Search, Navigate and Use Multimedia Museum Collections over the Web. In D. Bearman and J. Trant, editors, *Proceedings of Museums and the Web 2005*, pages 39–54, Vancouver, BC, Canada, April 13–17 2005.
- [3] M. Adriani and F. Arnely. Retrieving Images Using Cross-Language Text and Image Features. In C. Peters, F. C. Gey, J. Gonzalo, H. Müller, G. J. F. Jones, M. Kluck, B. Magnini, M. de Rijke, and D. Giampiccolo, editors, *Accessing Multilingual Information Repositories: Sixth Workshop of the Cross-Language Evaluation Forum (CLEF 2005)*, volume 4022 of *Lecture Notes in Computer Science (LNCS)*, pages 733–736, Vienna, Austria, September 20–22 2005. Springer.
- [4] R. Agrawal, T. Imielinski, and A. Swami. Mining Association Rules between Sets of Items in Large Databases. In *Proceedings of the 1993 ACM SIGMOD Conference*, pages 207–216, Washington, DC, USA, May 26–28 1993. ACM Press.
- [5] R. Agrawal and R. Srikant. Fast Algorithms for Mining Association Rules. In *Proceedings of the 20th Conference on Very Large Data Bases (VLDB)*, pages 487–499, Santiago, Chile, September 12–15 1994.

- [6] P. Aigrain, H. Zhang, and D. Petkovic. Content-based Representation and Retrieval of Visual Media: A State-of-the-Art Review. *Multimedia Tools and Applications*, 3(3):179–202, November 1996.
- [7] G. Amati, C. Carpineto, and G. Romano. Query difficulty, robustness and selective application of query expansion. In *Proceedings of the 25th European Conference on Information Retrieval (ECIR 2004)*, pages 127–137, Sunderland, UK, April 4–7 2004.
- [8] L. H. Armitage and P. G. B. Enser. Information Need in the Visual Document Domain. Technical Report Research and Innovation Report 27, British Library Research and Innovation Centre, London, UK, 1996.
- [9] L. H. Armitage and P. G. B. Enser. Analysis of user need in image archives. *Journal of Information Science*, 23(4):287–299, 1997.
- [10] T. Arndt. A Survey of Recent Research in Image Database Management. In *Proceedings of the 1990 IEEE Workshop on Visual Languages*, pages 92–97, Skokie, IL, USA, October 4–6 1990.
- [11] J. A. Aslam, V. Pavlu, and E. Yilmaz. A Statistical Method for System Evaluation Using Incomplete Judgments. In S. Dumais, E. N. Efthimiadis, D. Hawking, and K. Järvelin, editors, *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2006)*, pages 541–548, Seattle, WA, USA, August 6–11 2006. ACM Press.
- [12] E. Attalla and P. Siy. Robust Shape Similarity Retrieval Based on Contour Segmentation Polygonal Multiresolution and Elastic Matching. *Pattern Recognition*, 38(12):2229–2241, 2005.
- [13] I. A. A. Azzam. *Implicit Concept-Based Image Indexing and Retrieval for Visual Information Systems*. PhD thesis, School of Computer Science and Mathematics, Victoria University, Melbourne, Australia, 2006.

- [14] J. R. Bach, C. Fuller, A. Gupta, A. Hampapur, B. Horowitz, R. Humphrey, R. C. Jain, and C.-F. Shu. The Virage Image Search Engine: An Open Framework For Image Management. In I. K. Sethi and R. C. Jain, editors, *Storage and Retrieval for Image and Video Databases IV*, volume 2670 of *SPIE Proceedings*, pages 76–87, San Jose, CA, USA, March 1996.
- [15] A. Bagga and A. W. Biermann. Analyzing the Complexity of a Domain With Respect To An Information Extraction Task. In *Proceedings of The Tenth International Conference on Research on Computational Linguistics (ROCLING X)*, pages 175–194, Taipei, Taiwan, August 22–24 1997.
- [16] A. G. Balan, A. J. Traina, C. J. Traina, and P. M. Azevedo-Marques. Fractal Analysis of Image Textures for Indexing and Retrieval by Content. In *Proceedings of the 18th IEEE Symposium on Computer-Based Medical Systems (CBMS'05)*, pages 581–586, Dublin, Ireland, June 23–25 2005. IEEE Computer Society.
- [17] M. Barbieri, G. Mekenkamp, M. Ceccarelli, and J. Nesvadba. The Color Browser: A Content-Driven Linear Video Browsing Tool. In *Proceedings of the Second International Conference on Multimedia and Exposition (ICME'2001)*, pages 808–811, Tokyo, Japan, August 22–25 2001. IEEE Computer Society.
- [18] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. M. Blei, and M. I. Jordan. Matching Words and Pictures. *Journal of Machine Learning Research*, 3:1107–1135, March 2003.
- [19] K. Barnard, Q. Fan, R. Swaminathan, A. Hoogs, R. Collins, P. Rondot, and J. Kaufhold. Evaluation of Localized Semantics: Data, Methodology, and Experiments. Technical Report TR-05-08, University of Arizona, Computing Science, Tucson, AZ, USA, September 2005 (revised October 2006).
- [20] K. Barnard, L. Martin, B. Funt, and A. Coath. A data set for color research. *Color Research and Application*, 27(3):147–151, June 2002.

- [21] I. Bartolini, P. Ciaccia, and M. Patella. WARP: Accurate Retrieval of Shapes Using Phase of Fourier Descriptors and Time Warping Distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(1):142–147, January 2005.
- [22] G. E. Barton, R. C. Berwick, and E. S. Ristad. *Computational Complexity and Natural Language*. MIT Press, Cambridge, MA, USA, 1987.
- [23] N. Beckmann, H.-P. Kriegel, R. Schneider, and B. Seeger. The R*-Tree: An Efficient and Robust Access Method for Points and Rectangles. In H. Garcia-Molina and H. V. Jagadish, editors, *Proceedings of the 1990 ACM SIGMOD International Conference on Management of Data*, pages 322–331, Atlantic City, NJ, USA, May 23–25 1990. ACM Press.
- [24] S. Belongie, J. Malik, and J. Puzicha. Shape Matching and Object Recognition Using Shape Contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(12):509–522, April 2002.
- [25] A. B. Benitez, M. Beigi, and S.-F. Chang. Using Relevance Feedback in Content-Based Image Metasearch. *IEEE Internet Computing*, 2(4):59–69, 1998.
- [26] B. Berlin and P. Kay. *Basic Color Terms: Their Universality and Evolution*. University of California Press, Berkely, CA, USA, 1969.
- [27] R. Besançon and C. Millet. Data Fusion of Retrieval Results from Different Media: Experiments at ImageCLEF 2005. In C. Peters, F. C. Gey, J. Gonzalo, H. Müller, G. J. F. Jones, M. Kluck, B. Magnini, M. de Rijke, and D. Giampiccolo, editors, *Accessing Multilingual Information Repositories: Sixth Workshop of the Cross-Language Evaluation Forum (CLEF 2005)*, volume 4022 of *Lecture Notes in Computer Science (LNCS)*, pages 622–631, Vienna, Austria, September 20–22 2005. Springer.

- [28] R. Besançon and C. Millet. Using Text and Image Retrieval Systems. Lic2m experiments at ImageCLEF 2006. In *CLEF working notes*, Alicante, Spain, September 2006.
- [29] I. Biederman. Recognition-by-Components: A Theory of Human Image Understanding. *Psychological Review*, 94(2):115–147, 1987.
- [30] T. E. Bjoerge and E. Y. Chang. Why One Example Is Not Enough For An Image Query. In *Proceedings of the 2004 International Conference on Multimedia and Exposition (ICME'2004)*, pages 253–256, Taipei, Taiwan, June 27–30 2004. IEEE Computer Society.
- [31] A. Blaser, editor. *IBM Symposium: Data Base Techniques for Pictorial Applications*, volume 81 of *Lecture Notes in Computer Science (LNCS)*, Florence, Italy, June 20–22 1979. Springer.
- [32] M. Bober. MPEG-7 Visual Shape Descriptors. *IEEE Transactions on Circuits and Systems for Video Technology*, 11(6):716–719, June 2001.
- [33] C. Buckley. The SMART Lab Report: The Modern SMART Years (1980-1996). *SIGIR Forum*, 31(1):17–22, 1997.
- [34] C. Buckley and E. M. Voorhees. Evaluating evaluation measure stability. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2000)*, pages 33–40, Athens, Greece, July 24–28 2000. ACM Press.
- [35] C. Buckley and E. M. Voorhees. Retrieval evaluation with incomplete information. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2004)*, pages 25–32, Sheffield, UK, July 25–29 2004. ACM Press.

- [36] C. S. Candler, S. H. J. Uijtdehaage, and S. E. Dennis. Introducing HEAL: The Health Education Assets Library. *Academic Medicine*, 78(3):249–253, March 2003.
- [37] D. Carmel, E. Yom-Tov, A. Darlow, and D. Pelleg. What Makes a Query Difficult? In S. Dumais, E. N. Efthimiadis, D. Hawking, and K. Järvelin, editors, *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2006)*, pages 390–397, Seattle, WA, USA, August 6–11 2006. ACM Press.
- [38] D. Carmel, E. Yom-Tov, and I. Soboroff. SIGIR Workshop Report: Predicting Query Difficulty—Methods and Applications. *SIGIR Forum*, 39(2):25–28, 2005.
- [39] G. Carneiro and N. Vasconcelos. A database centric view of semantic image annotation and retrieval. In G. Marchionini, A. Moffat, J. Tait, R. Baeza-Yates, and N. Ziviani, editors, *Proceedings of the Twenty-Eighth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2005)*, pages 559–566, Salvador, Brazil, August 15–19 2005. ACM Press.
- [40] C. Carson, S. Belongie, H. Greenspan, and J. Malik. Region-Based Image Querying. In *IEEE Workshop on Content-based Access of Image and Video Libraries (CBAIVL'97)*, pages 42–51, San Juan, Puerto Rico, June 20 1997. IEEE Computer Society.
- [41] C. Carson, S. Belongie, H. Greenspan, and J. Malik. Color- and Texture-Based Image Segmentation Using EM and Its Application to Content-Based Image Retrieval. In *Proceedings of the Sixth International Conference on Computer Vision 1998 (ICCV-98)*, pages 675–682, Bombay, India, January 4–7 1998. Narosa Publishing House.

- [42] C. Carson, S. Belongie, H. Greenspan, and J. Malik. Image Segmentation Using Expectation-Maximization and Its Application to Image Querying. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(8):1026–1038, August 2002.
- [43] C. Carson, M. Thomas, S. Belongie, J. M. Hellerstein, and J. Malik. Blobworld: A System for Region-Based Image Indexing and Retrieval. In D. P. Huijsmans and A. W. M. Smeulders, editors, *Visual Information and Information Systems: Third International Conference (VISUAL'99)*, volume 1614 of *Lecture Notes in Computer Science (LNCS)*, pages 509–516, Amsterdam, The Netherlands, June 2–4 1999. Springer.
- [44] B. Carterette, J. Allan, and R. Sitaraman. Minimal Test Collections for Retrieval Evaluation. In S. Dumais, E. N. Efthimiadis, D. Hawking, and K. Järvelin, editors, *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2006)*, pages 268–275, Seattle, WA, USA, August 6–11 2006. ACM Press.
- [45] J. A. Catalán and J. S. Jin. Dimension Reduction of Texture Features for Image Retrieval Using Hybrid Associative Neural Networks. In *Proceedings of the 2000 IEEE International Conference on Multimedia and Exposition (ICME'2000)*, pages 1211–1214, New York City, NY, USA, July 30 - August 2 2000. IEEE Computer Society.
- [46] S. Chandrasekaran, B. S. Manjunath, Y. F. Wang, J. Winkeler, and H. Zhang. An Eigenspace Update Algorithm for Image Analysis. *Graphical Models and Image Processing*, 59(5):321–332, 1997.
- [47] E. Y. Chang and K.-T. Cheng. Supporting Subjective Image Queries without Seeding requirements – proposing test queries for Benchathlon. In G. Beretta and R. Schettini, editors, *Internet Imaging III*, volume 4672 of *SPIE Proceedings*, pages 225–232, San Jose, CA, USA, January 21–22 2002.

- [48] N.-S. Chang and K.-S. Fu. Query-by-Pictorial-Example. *IEEE Transactions on Software Engineering*, SE-6(6):519–524, November 1980.
- [49] S.-K. Chang and A. Hsu. Image Information Systems: Where Do We Go From Here? *IEEE Transactions on Knowledge and Data Engineering*, 4(5):431–442, October 1992.
- [50] S.-K. Chang and T. L. Kunii. Pictorial Data-Base Systems. *IEEE Computer*, 14(11):13–21, November 1981.
- [51] S.-K. Chang, Q. Y. Shi, and C. W. Yan. Iconic indexing by 2-D strings. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 9(3):413–428, 1987.
- [52] T. Chang and C.-C. J. Kuo. Texture Analysis and Classification with Tree-Structured Wavelet Transform. *IEEE Transactions on Image Processing*, 2(4):429–441, October 1993.
- [53] Y.-C. Chang and H.-H. Chen. Approaches Using a Word-Image Ontology and an Annotated Image Corpus as Intermedia for Cross-Language Image Retrieval. In C. Peters, P. D. Clough, F. C. Gey, J. Karlgren, B. Magnini, D. W. Oard, M. de Rijke, and M. Stempfhuber, editors, *Evaluation of Multilingual and Multi-modal Information Retrieval: Seventh Workshop of the Cross-Language Evaluation Forum (CLEF 2006)*, Lecture Notes in Computer Science (LNCS), Alicante, Spain, September 19–21 2006. Springer. (in press).
- [54] Y.-C. Chang, W.-C. L. Lin, and H.-H. Chen. A Corpus-Based Relevance Feedback Approach to Cross-Language Image Retrieval. In C. Peters, F. C. Gey, J. Gonzalo, H. Müller, G. J. F. Jones, M. Kluck, B. Magnini, M. de Rijke, and D. Giampiccolo, editors, *Accessing Multilingual Information Repositories: Sixth Workshop of the Cross-Language Evaluation Forum (CLEF 2005)*, volume 4022 of *Lecture Notes in Computer Science (LNCS)*, pages 592–601, Vienna, Austria, September 20–22 2005. Springer.

- [55] M. Charikar, C. Chekuri, T. Feder, and R. Motwani. Incremental Clustering and Dynamic Information Retrieval. In *Proceedings of the Twenty-Ninth Annual ACM Symposium on Theory of Computing (STOC '97)*, pages 626–635, El Paso, TX, USA, May 4–6 1997. ACM Press.
- [56] M. Chock, A. F. Cardenas, and A. Klinger. Database Structure and Manipulation Capabilities of the Picture Database Management System (PICDMS). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(4):484–492, July 1984.
- [57] G. Ciocca and R. Schettini. Content-based similarity retrieval of trademarks using relevance feedback. *Pattern Recognition*, 34(8):1639–1655, 2001.
- [58] C. W. Cleverdon. Report on the testing and analysis of an investigation into the comparative efficiency of indexing systems. Technical report, College of Aeronautics, Cranfield, UK, September 1962.
- [59] C. W. Cleverdon, L. Mills, and M. Keen. Factors Determining the Performance of Indexing Systems. Technical report, ASLIB Cranfield Research Project, Cranfield, UK, 1966.
- [60] P. D. Clough, M. Grubinger, T. Deselaers, A. Hanbury, and H. Müller. Overview of the ImageCLEF 2006 photographic retrieval and object annotation tasks. In C. Peters, P. D. Clough, F. C. Gey, J. Karlgren, B. Magnini, D. W. Oard, M. de Rijke, and M. Stempfhuber, editors, *Evaluation of Multilingual and Multi-modal Information Retrieval: Seventh Workshop of the Cross-Language Evaluation Forum (CLEF 2006)*, Lecture Notes in Computer Science (LNCS), Alicante, Spain, September 19–21 2006. Springer. (in press).
- [61] P. D. Clough, M. Grubinger, T. Deselaers, A. Hanbury, and H. Müller. Overview of the ImageCLEF 2006 photographic retrieval and object annotation tasks. In *CLEF working notes*, Alicante, Spain, September 20–22 2006.

- [62] P. D. Clough, H. Müller, T. Deselaers, M. Grubinger, T. M. Lehmann, J. Jensen, and W. Hersch. The CLEF 2005 Cross–Language Image Retrieval Track. In C. Peters, F. C. Gey, J. Gonzalo, H. Müller, G. J. F. Jones, M. Kluck, B. Magnini, M. de Rijke, and D. Giampiccolo, editors, *Accessing Multilingual Information Repositories: Sixth Workshop of the Cross-Language Evaluation Forum (CLEF 2005)*, volume 4022 of *Lecture Notes in Computer Science (LNCS)*, pages 535–557, Vienna, Austria, September 20–22 2005. Springer.
- [63] P. D. Clough, H. Müller, and M. Sanderson. Overview of the CLEF Cross–Language Image Retrieval Track (ImageCLEF) 2004. In C. Peters, P. D. Clough, J. Gonzalo, G. J. F. Jones, M. Kluck, and B. Magnini, editors, *Multilingual Information Access for Text, Speech and Images: Fifth Workshop of the Cross-Language Evaluation Forum (CLEF 2004)*, volume 3491 of *Lecture Notes in Computer Science (LNCS)*, pages 597–613, Bath, UK, September 15–17 2004. Springer.
- [64] P. D. Clough and M. Sanderson. The CLEF 2003 Cross Language Image Retrieval Track. In C. Peters, J. Gonzalo, M. Braschler, and M. Kluck, editors, *Comparative Evaluation of Multilingual Information Access Systems: Fourth Workshop of the Cross-Language Evaluation Forum (CLEF 2003)*, volume 3237 of *Lecture Notes in Computer Science (LNCS)*, pages 581–593, Trondheim, Norway, August 21–22 2003. Springer.
- [65] P. D. Clough, M. Sanderson, and H. Müller. The CLEF Cross Language Image Retrieval Track (ImageCLEF) 2004. In *Proceedings of the Third International Conference on Image and Video Retrieval (CIVR 2004)*, volume 3115 of *Lecture Notes in Computer Science (LNCS)*, pages 243–251, Dublin, Ireland, July 21–23 2004. Springer.
- [66] P. D. Clough, M. Sanderson, and N. Reid. The Eurovision St Andrews Collection of Photographs. *SIGIR Forum*, 40(1):21–30, 2006.

- [67] D. Comaniciu and P. Meer. Mean Shift: A Robust Approach Toward Feature Space Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619, May 2002.
- [68] G. V. Cormack, C. R. Palmer, and C. L. A. Clarke. Efficient Construction of Large Test Collections. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '98)*, pages 282–289, Melbourne, Australia, August 24–28 1998. ACM Press.
- [69] J. M. Corridoni, A. Del Bimbo, and E. Vicario. Image Retrieval by Color Semantics With Incomplete Knowledge. *Journal of the American Society for Information Science*, 49(3):267–282, 1998.
- [70] I. J. Cox, J. Ghosh, M. L. Miller, T. V. Papatomas, and P. N. Yianilos. Hidden Annotation in Content Based Image Retrieval. In *IEEE Workshop on Content-based Access of Image and Video Libraries (CBAIVL'99)*, pages 76–81, Fort Collins, CO, USA, June 22 1999. IEEE Computer Society.
- [71] I. J. Cox, M. L. Miller, T. P. Minka, T. V. Papatomas, and P. N. Yianilos. The Bayesian Image Retrieval System, PicHunter: Theory, Implementation, and Psychophysical Experiments. *IEEE Transactions on Image Processing*, 9(1):20–37, January 2000.
- [72] I. J. Cox, M. L. Miller, T. P. Minka, and P. N. Yianilos. An Optimized Interaction Strategy for Bayesian Relevance Feedback. In *Proceedings of the 1998 IEEE Conference on Computer Vision and Pattern Recognition (CVPR '98)*, pages 553–558, Santa Barbara, CA, USA, June 23–25 1998. IEEE Computer Society.
- [73] I. J. Cox, M. L. Miller, S. M. Omohundro, and P. N. Yianilos. Target Testing and the PicHunter Bayesian Multimedia Retrieval System. In *Advances in*

Digital Libraries (ADL'96), pages 66–75, Library of Congress, Washington, DC, USA, May 13–15 1996.

- [74] S. Cronen-Townsend, Y. Zhou, and W. B. Croft. Predicting query performance. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2002)*, pages 299–306, Tampere, Finland, August 11–15 2001. ACM Press.
- [75] J. F. Cullen, J. J. Hull, and P. E. Hart. Document Image Database Retrieval and Browsing Using Texture Analysis. In *Proceedings of the 4th International Conference on Document Analysis and Recognition (ICDAR '97)*, volume 2, pages 718–721, Ulm, Germany, August 18–20 1997. IEEE Computer Society.
- [76] P. Curtioni, L. Dini, and V. D. Tomaso. CELI Participation at ImageCLEF 2006: Comparison with the Ad-hoc Track. In *CLEF working notes*, Alicante, Spain, September 2006.
- [77] R. Datta, J. Li, and J. Z. Wang. Content-based image retrieval: approaches and trends of the new age. In *Proceedings of the Seventh ACM SIGMM International Workshop on Multimedia Information Retrieval (MIR'05)*, pages 253–262, Singapore, November 10–11 2005. ACM Press.
- [78] I. Daubechies. The Wavelet Transform, Time-Frequency Localization and Signal Analysis. *IEEE Transactions on Information Theory*, 36(5):961–1005, September 1990.
- [79] J. G. Daugman. High Confidence Visual Recognition of Persons By a Test of Statistical Independence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(11):1148–1161, 1993.
- [80] N. Day. MPEG-7: Daring to Describe Multimedia Content. *XML-Journal*, 1(6):24–27, 2000.

- [81] A. P. de Vries, M. G. L. M. van Doorn, H. M. Blanken, and P. M. G. Apers. The Mirror MMDBMS architecture. In M. P. Atkinson, M. E. Orłowska, P. Valduriez, S. B. Zdonik, and M. L. Brodie, editors, *Proceedings of 25th International Conference on Very Large Data Bases (VLDB'99)*, pages 758–761, Edinburgh, Scotland, September 7–11 1999. Morgan Kaufmann.
- [82] A. del Bimbo. *Visual Information Retrieval*. Academic Press, 1999.
- [83] Y. Deng, B. S. Manjunath, C. S. Kenney, M. S. Moore, and H. Shin. An Efficient Color Representation for Image Retrieval. *IEEE Transactions on Image Processing*, 10(1):140–147, 2001.
- [84] L. Denoyer and P. Gallinari. The Wikipedia XML Corpus. *SIGIR Forum*, 40(1):64–69, June 2006.
- [85] T. Deselaers. Features for Image Retrieval. Master's thesis, Human Language Technology and Pattern Recognition Group, RWTH Aachen University, Aachen, Germany, December 2003.
- [86] T. Deselaers, H. Müller, P. Clough, H. Ney, and T. M. Lehmann. The CLEF 2005 Automatic Medical Image Annotation Task. *International Journal of Computer Vision*, 2007 (to appear).
- [87] T. Deselaers, T. Weyand, D. Keysers, W. Macherey, and H. Ney. FIRE in ImageCLEF 2005: Combining Content-based Image Retrieval with Textual Information Retrieval. In C. Peters, F. C. Gey, J. Gonzalo, H. Müller, G. J. F. Jones, M. Kluck, B. Magnini, M. de Rijke, and D. Giampiccolo, editors, *Accessing Multilingual Information Repositories: Sixth Workshop of the Cross-Language Evaluation Forum (CLEF 2005)*, volume 4022 of *Lecture Notes in Computer Science (LNCS)*, pages 652–661, Vienna, Austria, September 20–22 2005. Springer.
- [88] T. Deselaers, T. Weyand, and H. Ney. Image Retrieval and Annotation Using Maximum Entropy. In C. Peters, P. D. Clough, F. C. Gey, J. Karlgren,

- B. Magnini, D. W. Oard, M. de Rijke, and M. Stempfhuber, editors, *Evaluation of Multilingual and Multi-modal Information Retrieval: Seventh Workshop of the Cross-Language Evaluation Forum (CLEF 2006)*, Lecture Notes in Computer Science (LNCS), Alicante, Spain, September 19–21 2006. Springer. (in press).
- [89] M. R. Díaz-Galiano, M. A. García-Cumbreras, M. T. Martín-Valdivia, A. Montejo-Raez, and L. A. Ureña López. Using Information Gain to Improve the ImageCLEF 2006 Collection. In C. Peters, P. D. Clough, F. C. Gey, J. Karlgren, B. Magnini, D. W. Oard, M. de Rijke, and M. Stempfhuber, editors, *Evaluation of Multilingual and Multi-modal Information Retrieval: Seventh Workshop of the Cross-Language Evaluation Forum (CLEF 2006)*, Lecture Notes in Computer Science (LNCS), Alicante, Spain, September 19–21 2006. Springer. (in press).
- [90] A. Dimai. Assessment of Effectiveness of Content-Based Image Retrieval Systems. In D. P. Huijsmans and A. W. M. Smeulders, editors, *Visual Information and Information Systems: Third International Conference (VISUAL'99)*, volume 1614 of *Lecture Notes in Computer Science (LNCS)*, pages 525–532, Amsterdam, The Netherlands, June 2–4 1999. Springer.
- [91] M. H. Do and M. Vetterli. Wavelet-Based Texture Retrieval Using Generalized Gaussian Density and Kullback-Leibler Distance. *IEEE Transactions on Image Processing*, 11(2):146–158, February 2002.
- [92] P. Dollár, Z. Tu, and S. Belongie. Supervised Learning of Edges and Object Boundaries. In *Proceedings of the 2006 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2006)*, pages 1964–1971, New York, NY, USA, June 17–22 2006. IEEE Computer Society.
- [93] K. M. Donald and G. J. F. Jones. Dublin City University at CLEF 2006: Experiments for the ImageCLEF Photo Collection Standard Ad Hoc Task. In *CLEF working notes*, Alicante, Spain, September 2006.

- [94] M. D. Dunlop. Reflections on MIRA: Interactive Evaluation in Information Retrieval. *Journal of the American Society for Information Science*, 51(14):1269–1274, 2000.
- [95] D. Dupplaw, S. Dasmahapatra, B. Hu, P. Lewis, and N. Shadbolt. Multimedia Distributed Knowledge Management in MIAKT. In *Proceedings of the ISWC 2004 Workshop on Knowledge Markup and Semantic Annotation*, pages 81–90, Hiroshima, Japan, November 8 2004.
- [96] P. Duygulu, K. Barnard, J. F. G. de Freitas, and D. A. Forsyth. Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary. In *Proceedings of the 7th European Conference on Computer Vision-Part IV (ECCV '02)*, volume 2353 of *Lecture Notes in Computer Science (LNCS)*, pages 97–112, Copenhagen, Denmark, May 28–31 2002. Springer.
- [97] J. P. Eakins, J. M. Boardman, and M. E. Graham. Similarity Retrieval of Trademark Images. *IEEE Multimedia*, 5(2):53–63, April – June 1998.
- [98] J. P. Eakins, J. M. Boardman, and K. Shields. Retrieval of Trade Mark Images by Shape Feature – the ARTISAN Project. In *IEE Colloquium on Intelligent Image Databases*, pages 9/1 – 9/6, London, UK, May 22 1996.
- [99] J. P. Eakins and M. E. Graham. Content-Based Image Retrieval. Technical Report JTAP-039, JISC Technology Application Program, University of Northumbria, Newcastle upon Tyne, UK, 2000.
- [100] R. Egas, D. P. Huijsmans, M. S. Lew, and N. Sebe. Adapting k-d Trees to Visual Retrieval. In D. P. Huijsmans and A. W. M. Smeulders, editors, *Visual Information and Information Systems: Third International Conference (VISUAL'99)*, volume 1614 of *Lecture Notes in Computer Science (LNCS)*, pages 533–540, Amsterdam, The Netherlands, June 2–4 1999. Springer.

- [101] M. J. Egenhofer. Spatial-Query-by-Sketch. In *Proceedings of the IEEE Symposium on Visual Languages (VL 1996)*, pages 6–67, Boulder, CO, USA, September 3–6 1996.
- [102] K. Eguchi, K. Kuriyama, and N. Kando. Sensitivity of IR Systems Evaluation to Topic Difficulty. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, volume 2, pages 585–589, Las Palmas de Gran Canaria, Spain, May 29–31 2002.
- [103] P. G. B. Enser. Query analysis in a visual information retrieval context. *Journal of Document and Text Management*, 1(1):25–52, 1993.
- [104] P. G. B. Enser. Pictorial Information Retrieval. *Journal of Documentation*, 51(2):126–170, 1995.
- [105] P. G. B. Enser. Visual Information Retrieval: Seeking the Alliance of Concept-Based and Content-Based Paradigms. *Journal of Information Science*, 26(4):199–210, 2000.
- [106] P. G. B. Enser and C. G. McGregor. Analysis of Visual Information Retrieval Queries. Technical Report Research and Development Report 6104, British Library, London, UK, 1992.
- [107] P. G. B. Enser, C. J. Sandom, and P. H. Lewis. Surveying the Reality of Semantic Image Retrieval. In S. Bres and R. Laurini, editors, *8th International Conference On Visual Information and Information Systems (VIS'2005)*, volume 3736 of *Lecture Notes in Computer Science (LNCS)*, pages 177–188, Amsterdam, The Netherlands, July 5 2006. Springer.
- [108] D. A. Evans, J. G. Shanahan, and V. Sheftel. Topic structure modeling. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2002)*, pages 417–418, Tampere, Finland, August 11–15 2001. ACM Press.

- [109] M. Everingham, A. Zisserman, C. K. I. Williams, and L. van Gool. The PASCAL Visual Object Classes Challenge 2006 (VOC2006) Results. Technical report, University of Oxford, Oxford, UK, September 11 2006.
- [110] M. Everingham, A. Zisserman, C. K. I. Williams, L. van Gool, M. Allan, C. M. Bishop, O. Chapelle, N. Dalal, T. Deselaers, G. Dorko, S. Duffner, J. Eichhorn, J. D. R. Farquhar, M. Fritz, C. Garcia, T. Griffiths, F. Jurie, D. Keysers, M. Koskela, J. Laaksonen, D. Larlus, B. Leibe, H. Meng, H. Ney, B. Schiele, C. Schmid, E. Seemann, J. Shawe-Taylor, A. Storkey, S. Szedmak, B. Triggs, I. Ulusoy, V. Viitaniemi, and J. Zhang. The 2005 PASCAL Visual Object Classes Challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Textual Entailment (PASCAL Workshop 05)*, number 3944 in Lecture Notes in Artificial Intelligence, pages 117–176, Southampton, UK, 2006. Springer.
- [111] C. Faloutsos, R. Barber, M. Flickner, J. Hafner, W. Niblack, D. Petkovic, and W. Equitz. Efficient and Effective Querying by Image Content. *Journal of Intelligent Information Systems*, 3(3-4):231–262, 1994.
- [112] C. Faloutsos and K.-I. Lin. FastMap: A Fast Algorithm for Indexing, Data-Mining and Visualization of Traditional and Multimedia Datasets. In *Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data (SIGMOD '95)*, pages 163–174, San Jose, CA, USA, May 22–25 1995. ACM Press.
- [113] C. Fellbaum. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA, USA, 1998.
- [114] D. D. Feng, W.-C. Siu, and H. Zhang, editors. *Multimedia Information Retrieval and Management: Technological Fundamentals and Applications*. Springer, 2003.

- [115] S. L. Feng, R. Manmatha, and V. Lavrenko. Multiple Bernoulli Relevance Models for Image and Video Annotation. In *Proceedings of the 2004 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2004)*, pages 1002–1009, Washington, DC, USA, 27 June – 2 July 2004. IEEE Computer Society.
- [116] S. Flank. Sentences vs. Phrases: Syntactic Complexity in Multimedia Information Retrieval. In *Workshop on Syntactic and Semantic Complexity in Natural Language Processing Systems (NAACL-ANLP 2000)*, pages 1–5, Seattle, WA, USA, 2000. Association for Computational Linguistics.
- [117] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker. Query by Image and Video Content: The QBIC System. *IEEE Computer*, 28(9):23–32, September 1995.
- [118] C. Fluhr, P.-A. Moëllic, and P. Hede. ImageEVAL: Usage-oriented multimedia information retrieval evaluation. In A. Hanbury, H. Müller, and P. D. Clough, editors, *Proceedings of the Second Workshop on Image and Video Retrieval Evaluation*, pages 3–8, Alicante, Spain, September 19 2006. MUSCLE Network of Excellence.
- [119] J. D. Foley, A. van Dam, S. K. Feiner, and J. F. Hughes. *Computer Graphics: Principles and Practice (2nd edition)*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1990.
- [120] D. A. Forsyth. Benchmarks for storage and retrieval in multimedia databases. In G. Beretta and R. Schettini, editors, *Internet Imaging III*, volume 4672 of *SPIE Proceedings*, pages 240–247, San Jose, CA, USA, January 21–22 2002.
- [121] D. A. Forsyth and M. M. Fleck. Automatic Detection of Human Nudes. *International Journal of Computer Vision*, 32(1):63–77, 1999.

- [122] E. Fox. Characteristics of Two New Experimental Collections in Computer and Information Science Containing Textual and Bibliographic Concepts. Technical report, Number 83-561, Computing Science Department, Cornell University, Ithaca, NY, USA, September 1983.
- [123] J. M. Francos, A. Z. Meiri, and B. Porat. A Unified Texture Model Based on a 2-D Wold-Like Decomposition. *IEEE Transactions on Signal Processing*, 41(8):2665–2678, August 1993.
- [124] C. Frankel, M. J. Swain, and V. Athitsos. WebSeer: An Image Search Engine for the World Wide Web. Technical Report 96-14, University of Chicago, Chicago, IL, USA, August 1996.
- [125] C. O. Frost, B. Taylor, A. Noakes, S. Markel, D. Torres, and K. M. Drabentstott. Browse and Search Patterns in a Digital Image Database. *Information Retrieval*, 1(4):287–313, January 2000.
- [126] U. Gargi and R. Kasturi. Image Database Querying Using a Multi-Scale Localized Color Representation. In *IEEE Workshop on Content-based Access of Image and Video Libraries (CBAIVL'99)*, pages 28–32, Fort Collins, CO, USA, June 22 1999. IEEE Computer Society.
- [127] J.-M. Geusebroek, G. J. Burghouts, and A. W. M. Smeulders. The Amsterdam Library of Object Images. *International Journal of Computer Vision*, 61(1):103–112, 2005.
- [128] J.-M. Geusebroek, R. van den Boogaard, A. W. M. Smeulders, and H. Geerts. Color Invariance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(12):1338–1350, December 2001.
- [129] T. Gevers and A. W. M. Smeulders. A Content-Based Image Search System for the World Wide Web. In *Proceedings of the Second International Conference On Visual Information Systems (VISUAL'97)*, pages 96–100, San Diego, CA, USA, December 15–17 1997. Knowledge Systems Institute.

- [130] T. Gevers and A. W. M. Smeulders. PicToSeek: Combining Color and Shape Invariant Feature for Image Retrieval. *IEEE Transactions on Image Processing*, 9(1):102–119, January 2000.
- [131] A. Ghoshal, P. Ircing, and S. Khudanpur. Hidden Markov models for automatic annotation and content-based retrieval of images and video. In G. Marchionini, A. Moffat, J. Tait, R. Baeza-Yates, and N. Ziviani, editors, *Proceedings of the Twenty-Eighth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2005)*, pages 544–551, Salvador, Brazil, August 15–19 2005. ACM Press.
- [132] H. Gilbert and K. S. Jones. Statistical bases of relevance assessment for the ‘ideal’ information retrieval test collection. British Library Research and Development Report 5481, Computer Laboratory, University of Cambridge, 1979.
- [133] G. L. Gimel’farb and A. K. Jain. On Retrieving Textured Images From an Image Database. *Pattern Recognition*, 29(9):1461–1483, 1996.
- [134] K. Glatz-Krieger, D. Glatz, M. Gysel, M. Dittler, and M. J. Mihatsch. Web-basierte Lernwerkzeuge für die Pathologie. *Der Pathologe*, 24(5):394–399, 2003.
- [135] R. C. Gonzales and P. Wintz. *Digital Image Processing*. Addison-Wesley, Reading, MA, USA, 1987.
- [136] J. Gonzalo, P. D. Clough, and A. Vallin. Overview of the CLEF 2005 Interactive Track. In C. Peters, F. C. Gey, J. Gonzalo, H. Müller, G. J. F. Jones, M. Kluck, B. Magnini, M. de Rijke, and D. Giampiccolo, editors, *Accessing Multilingual Information Repositories: Sixth Workshop of the Cross-Language Evaluation Forum (CLEF 2005)*, volume 4022 of *Lecture Notes in Computer Science (LNCS)*, pages 251–262, Vienna, Austria, September 20–22 2005. Springer.

- [137] J. Gonzalo, J. Karlgren, and P. D. Clough. iCLEF 2006 Overview: Searching the FlickrR WWW photo-sharing repository. In C. Peters, P. D. Clough, F. C. Gey, J. Karlgren, B. Magnini, D. W. Oard, M. de Rijke, and M. Stempfhuber, editors, *Evaluation of Multilingual and Multi-modal Information Retrieval: Seventh Workshop of the Cross-Language Evaluation Forum (CLEF 2006)*, Lecture Notes in Computer Science (LNCS), Alicante, Spain, September 19–21 2006. Springer. (in press).
- [138] A. Goodrum and A. Spink. Visual Information Seeking: A Study of Image Queries on the World Wide Web. In *Proceedings of the 62nd Annual Meeting of the American Society for Information Science (ASIS 1999)*, pages 665–674, Washington, DC, USA, 31 October – 4 November 1999.
- [139] A. A. Goodrum. Image Information Retrieval: An Overview of Current Research. *Informing Science, Special Issue on Information Science Research*, 3(2):63–66, 2000.
- [140] C. C. Gotlieb and H. E. Kreyzig. Texture Descriptors Based on Co-occurrence Matrices. *Computer Vision, Graphics, and Image Processing*, 51(1):70–86, 1990.
- [141] V. Gouet and N. Boujemaa. On the Robustness of Color Points of Interest for Image Retrieval. In *Proceedings of the 2002 IEEE International Conference on Image Processing (ICIP 2002)*, pages 377–380, Rochester, NY, USA, September 22–25 2002. IEEE Computer Society.
- [142] J. Gray, editor. *The Benchmark Handbook for Database and Transaction Systems (2nd Edition)*. Morgan Kaufmann, 1993.
- [143] C. Grigorescu and N. Petkov. Distance Sets for Shape Filters and Shape Recognition. *IEEE Transactions on Image Processing*, 12(10):1274–1286, October 2003.

- [144] M. Grubinger. SCM6102: Research Project Report: Benchmarking for Content Based Visual Information Search. Technical report, Victoria University of Technology, Melbourne, Australia, November 2002.
- [145] M. Grubinger, P. D. Clough, and C. H. C. Leung. The IAPR TC-12 Benchmark for Visual Information Search. *IAPR Newsletter April 2006*, 28(2):10–12, 2006.
- [146] M. Grubinger, P. D. Clough, H. Müller, and T. Deselears. The IAPR-TC12 Benchmark: A New Evaluation Resource for Visual Information Systems. In *International Workshop OntoImage'2006 Language Resources for Content-Based Image Retrieval, held in conjunction with LREC'06*, pages 13–23, Genoa, Italy, May 22 2006.
- [147] M. Grubinger and C. H. C. Leung. A Benchmark for Performance Calibration in Visual Information Search. In H. Ip, editor, *The 2003 International Conference on Visual Information Systems (VIS'2003)*, pages 414–419, Miami, FL, USA, September 2003. Knowledge Systems Institute.
- [148] M. Grubinger and C. H. C. Leung. Incremental Benchmark Development and Administration. In J. K. Wu and C. H. C. Leung, editors, *The Seventh International Conference on Visual Information Systems (VIS'2004)*, pages 328–333, San Francisco, CA, USA, September 2004. Knowledge Systems Institute.
- [149] M. Grubinger, C. H. C. Leung, and P. D. Clough. Linguistic Estimation of Topic Difficulty in Cross-Language Image Retrieval. In C. Peters, F. C. Gey, J. Gonzalo, H. Müller, G. J. F. Jones, M. Kluck, B. Magnini, M. de Rijke, and D. Giampiccolo, editors, *Assessing Multilingual Information Repositories: Sixth Workshop of the Cross-Language Evaluation Forum (CLEF 2005)*, volume 4022 of *Lecture Notes in Computer Science (LNCS)*, pages 558–566, Vienna, Austria, September 20–22 2005. Springer.

- [150] M. Grubinger, C. H. C. Leung, and P. D. Clough. The IAPR Benchmark for Assessing Image Retrieval Performance in Cross-Language Evaluation Tasks. In A. Hanbury and H. Müller, editors, *MUSCLE/ImageCLEF Workshop on Image and Video Retrieval Evaluation*, pages 33–50, Vienna, Austria, September 20 2005. MUSCLE Network of Excellence.
- [151] V. N. Gudivada and V. V. Raghavan. Content-Based Image Retrieval Systems. *IEEE Computer*, 28(9):18–22, September 1995.
- [152] V. N. Gudivada and V. V. Raghavan. Design and Evaluation of Algorithms for Image Retrieval by Spatial Similarity. *ACM Transactions on Information Systems*, 13(2):115–144, 1995.
- [153] V. N. Gudivada, V. V. Raghavan, W. I. Grosky, and R. Kasanagottu. Information Retrieval on the World Wide Web. *IEEE Internet Computing*, 1(5):58–68, September 1997.
- [154] N. J. Gunther and G. Beretta. A Benchmark for Image Retrieval using Distributed Systems over the Internet: BIRDS-I. Technical Report HPL-2000-162, HP Labs, Palo Alto, San Jose, CA, USA, 2001.
- [155] F. Guo, J. Jin, and D. Feng. Measuring Image Similarity Using the Geometrical Distribution of Image Contents. In *Proceedings of the 1998 Fourth International Conference on Signal Processing (ICSP '98)*, volume 2, pages 1108–1112, Beijing, China, December 12–16 1998.
- [156] A. Guttman. R-Trees: A Dynamic Index Structure for Spatial Searching. In *Proceedings of the 1984 ACM SIGMOD International Conference on Management of Data (SIGMOD '84)*, pages 47–57, Boston, MA, USA, June 18–21 1984. ACM Press.
- [157] E. Hadjidemetriou, M. D. Grossberg, and S. K. Nayar. Multiresolution Histograms and Their Use for Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(7):831–847, July 2004.

- [158] J. Hafner, H. S. Sawhney, W. Equitz, M. Flickner, and W. Niblack. Efficient Color Histogram Indexing for Quadratic Form Distance Functions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(7):729–736, 1995.
- [159] A. Hampapur, A. Gupta, B. Horowitz, C.-F. Shu, C. Fuller, J. Bach, M. Gorkani, and R. C. Jain. Virage Video Engine. In I. K. Sethi and R. C. Jain, editors, *Storage and Retrieval for Image and Video Databases V*, volume 3022 of *SPIE Proceedings*, pages 188–198, San Jose, CA, USA, February 1997.
- [160] A. Hanbury. A Dataset of Annotated Animals. In A. Hanbury, H. Müller, and P. D. Clough, editors, *Proceedings of the Second Workshop on Image and Video Retrieval Evaluation*, pages 19–27, Alicante, Spain, September 19 2006. MUSCLE Network of Excellence.
- [161] A. Hanbury. MUSCLE Guide to Annotation. Technical Report Version 2.12, Pattern Recognition and Image Processing Group, Vienna University of Technology, Austria, 2006.
- [162] J. S. Hare. *Saliency for Image Description and Retrieval*. PhD thesis, School of Electronics and Computer Science, University of Southampton, Southampton, UK, April 2006.
- [163] J. S. Hare and P. H. Lewis. Saliency-based Models of Image Content and their Application to Auto-Annotation by Semantic Propagation. In *Proceedings of the Second European Semantic Web Conference (ESWC 2005)*, Heraklion, Crete, Greece, May 29 - June 1 2005.
- [164] J. S. Hare, P. H. Lewis, P. G. B. Enser, and C. J. Sandom. Mind the Gap: Another Look at the Problem of the Semantic Gap in Image Retrieval. In E. Y. Chang, A. Hanjalic, and N. Sebe, editors, *Proceedings of Multimedia Content Analysis, Management and Retrieval 2006*, volume 6073 of *SPIE Proceedings*, pages 607309–1, San Jose, CA, USA, January 17–19 2006.

- [165] D. K. Harman. Overview of the First Text REtrieval Conference (TREC-1). In *The Text REtrieval Conference (TREC-1)*, pages 1–20, Gaithersburg, MD, USA, November 4–6 1992. Department of Commerce, National Institute of Standards and Technology.
- [166] D. K. Harman. Overview of the Second Text REtrieval Conference (TREC-2). In *The Second Text REtrieval Conference (TREC-2)*, pages 1–20, Gaithersburg, MD, USA, 31 August – 2 September 1993. Department of Commerce, National Institute of Standards and Technology.
- [167] D. K. Harman. Overview of the Fourth Text REtrieval Conference (TREC-4). In D. K. Harman, editor, *Proceedings of the Fourth Text REtrieval Conference (TREC-4)*, pages 1–23, Gaithersburg, MD, USA, November 1–3 1995. Department of Commerce, National Institute of Standards and Technology.
- [168] D. K. Harman and C. Buckley. The NRRC reliable information access (RIA) workshop. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2004)*, pages 528–529, Sheffield, UK, July 25–29 2004. ACM Press.
- [169] S. K. Hastings. Query Categories in a Study of Intellectual Access to Digitized Art Images. In T. Kinney, editor, *Proceedings of the 58th ASIS Annual Meeting*, pages 3–8, Chicago, IL, USA, October 9–12 1995. Information Today, Inc.
- [170] X. He. Incremental Semi-Supervised Subspace Learning for Image Retrieval. In *Proceedings of the 12th ACM International Conference on Multimedia (ACM Multimedia 2004)*, pages 2–8, New York, NY, USA, October 10–16 2004. ACM Press.
- [171] S. C.-H. Hoi and M. R. Lyu. A Novel Log-based Relevance Feedback Technique in Content-based Image Retrieval. In *Proceedings of the 12th ACM*

International Conference on Multimedia (ACM Multimedia 2004), pages 24–31, New York, NY, USA, October 10–16 2004. ACM Press.

- [172] S. C.-H. Hoi, J. Zhu, and M. R. Lyu. CUHK at ImageCLEF 2005: Cross-Language and Cross-Media Image Retrieval. In C. Peters, F. C. Gey, J. Gonzalo, H. Müller, G. J. F. Jones, M. Kluck, B. Magnini, M. de Rijke, and D. Giampiccolo, editors, *Accessing Multilingual Information Repositories: Sixth Workshop of the Cross-Language Evaluation Forum (CLEF 2005)*, volume 4022 of *Lecture Notes in Computer Science (LNCS)*, pages 602–611, Vienna, Austria, September 20–22 2005. Springer.
- [173] L. Hollink, G. Nguyen, A. T. G. Schreiber, J. Wielemaker, B. J. Wielinga, and M. Worring. Adding Spatial Semantics to Image Annotations. In P. Cimiano, F. Ciravegna, E. Motta, and V. Uren, editors, *Application of Language and Semantic Technologies to support Knowledge Management Processes (LSTKM 2004)*, Whittlebury Hall, Northamptonshire, UK, October 2004.
- [174] L. Hollink, A. T. G. Schreiber, J. Wielemaker, and B. J. Wielinga. Semantic Annotation of Image Collections. In S. Handschuh, M. Koivunen, R. Dieng, and S. Staab, editors, *Knowledge Capture 2003 – Proceedings Knowledge Markup and Semantic Annotation Workshop*, pages 41–48, Sanibel, FL, USA, October 25–26 2003.
- [175] L.-J. Hove. Extending Image Retrieval Systems with a Thesaurus for Shapes. Master’s thesis, University of Bergen, Norway, October 2004.
- [176] B. Hu, S. Dasmahapatra, P. Lewis, and N. Shadbolt. Ontology-based medical image annotation with description logics. In *Proceedings of the 15th IEEE International Conference on Tools with Artificial Intelligence (ICTAI ’03)*, pages 77–82, Sacramento, CA, USA, November 3–5 2003. IEEE Computer Society.

- [177] M.-K. Hu. Visual Pattern Recognition by Moment Invariants. *IRE Transactions on Information Theory*, 8(2):179–187, February 1962.
- [178] J. Huang, S. R. Kumar, M. Mitra, W.-J. Zhu, and R. Zabih. Image Indexing Using Color Correlograms. In *Proceedings of the 1997 IEEE Conference on Computer Vision and Pattern Recognition (CVPR '97)*, pages 762–768, San Juan, Puerto Rico, June 17–19 1997. IEEE Computer Society.
- [179] J. Hunter. Adding Multimedia to the Semantic Web - Building an MPEG-7 Ontology. In *Proceedings of the First Semantic Web Working Symposium (SWWS)*, pages 261–283, Stanford, CA, USA, July 30 - August 1 2001.
- [180] D. P. Huttenlocher, G. A. Klandermann, and W. J. Rucklidge. Comparing Images Using the Hausdorff Distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(9):850–863, September 1993.
- [181] D. F. Huynh, S. M. Drucker, P. Baudisch, and C. Wong. Time Quilt: Scaling Up Zoomable Photo Browsers For Large, Unstructured Photo Collections. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '05)*, pages 1937–1940, Portland, OR, USA, April 2–7 2005. ACM Press.
- [182] W.-S. Hwang, J. J. Weng, M. Fang, and J. Qian. A Fast Image Retrieval Algorithm with Automatically Extracted Discriminant Features. In *IEEE Workshop on Content-based Access of Image and Video Libraries (CBAIVL'99)*, pages 8–12, Fort Collins, CO, USA, June 22 1999. IEEE Computer Society.
- [183] E. Hyvönen, A. Styrman, and S. Saarela. Ontology-based Image Retrieval. In *Proceedings of the XML Finland 2002 Conference: Toward the Semantic Web and Web Services*, pages 15–27, Helsinki, Finland, October 21–22 2002. HIIT Publication.
- [184] M. Inoue. Using Visual Linkages for Multilingual Image Retrieval. In C. Peters, P. D. Clough, F. C. Gey, J. Karlgren, B. Magnini, D. Oard, M. de Rijke,

- and M. Stempfhuber, editors, *Evaluation of Multilingual and Multi-modal Information Retrieval: Seventh Workshop of the Cross-Language Evaluation Forum (CLEF 2006)*, Lecture Notes in Computer Science (LNCS), Alicante, Spain, September 19–21 2006. Springer. (in press).
- [185] H. Ip, editor. *The 2003 International Conference on Visual Information Systems (VIS'2003)*, Miami, FL, USA, September 2003. Knowledge Systems Institute.
- [186] T. Ishioka. Evaluation of Criteria for Information Retrieval. In *Web Intelligence*, pages 425–431, Beijing, China, October 13–17 2003. IEEE Computer Society.
- [187] A. K. Jain and A. Vailaya. Image Retrieval Using Color and Shape. *Pattern Recognition*, 29(8):1233–1244, August 1996.
- [188] A. K. Jain and A. Vailaya. Shape-Based Retrieval: A Case Study With Trademark Images. *Pattern Recognition*, 31(9):1369–1390, 1998.
- [189] R. C. Jain. World-Wide Maze. *IEEE MultiMedia*, 2(2):3–6, 1995.
- [190] A. C. Jalba, M. H. F. Wilkinson, and J. B. T. M. Roerdink. Shape Representation and Recognition Through Morphological Curvature Scale Spaces. *IEEE Transactions on Image Processing*, 15(2):331–341, February 2006.
- [191] F. P. Janecek. *Interactive Semantic Fisheye Views For Information Workspaces*. PhD thesis, École Polytechnique Fédérale de Lausanne, Section d'Informatique, Lausanne, Switzerland, 2004.
- [192] B. J. Jansen, A. Spink, and J. Pedersen. An Analysis of Multimedia Searching on AltaVista. In *Proceedings of the 5th ACM SIGMM International Workshop on Multimedia Information Retrieval (MIR'03)*, pages 186–192, Berkeley, CA, USA, November 7 2003. ACM Press.

- [193] J. Jeon, V. Lavrenko, and R. Manmatha. Automatic Image Annotation and Retrieval using Cross-Media Relevance Models. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2003)*, pages 119–126, Toronto, Canada, July 28 – August 01 2003. ACM Press.
- [194] J. Jeon and R. Manmatha. Using Maximum Entropy for Automatic Image Annotation. In *Proceedings of the Third International Conference on Image and Video Retrieval (CIVR 2004)*, volume 3115 of *Lecture Notes in Computer Science (LNCS)*, pages 24–32, Dublin, Ireland, July 21–23 2004. Springer.
- [195] J. S. Jin, R. Kurniawati, G. Xu, and X. Bai. Using Browsing to Improve Content-Based Image Retrieval. In C.-C. J. Kuo, S.-F. Chang, and S. Panchanathan, editors, *Multimedia Storage and Archiving Systems III*, volume 3527 of *SPIE Proceedings*, pages 101–109, San Jose, CA, USA, October 1998.
- [196] F. Jing, B. Zhang, F. Lin, W.-Y. Ma, and H.-J. Zhang. A novel region-based Image Retrieval Method Using Relevance Feedback. In *Proceedings of the ACM Multimedia Workshop on Multimedia Information Retrieval (MIR 2001)*, pages 28–31, Ottawa, Canada, October 5 2001. ACM Press.
- [197] G. J. F. Jones and K. McDonald. Dublin City University at CLEF 2005: Experiments with the ImageCLEF St Andrew’s Collection. In C. Peters, F. C. Gey, J. Gonzalo, H. Müller, G. J. F. Jones, M. Kluck, B. Magnini, M. de Rijke, and D. Giampiccolo, editors, *Accessing Multilingual Information Repositories: Sixth Workshop of the Cross-Language Evaluation Forum (CLEF 2005)*, volume 4022 of *Lecture Notes in Computer Science (LNCS)*, pages 567–573, Vienna, Austria, September 20–22 2005. Springer.
- [198] C. Jörgensen. The Applicability of Existing Classification Systems to Image Attributes: A Selected Review. *Knowledge Organisation and Change*, 5:189–197, 1996.

- [199] C. Jörgensen. Retrieving the Unretrievable in Electronic Imaging Systems: Emotions, Themes and Stories. In B. Rogowitz and T. N. Pappas, editors, *Human Vision and Electronic Imaging IV*, volume 3644 of *SPIE Proceedings*, pages 348–355, San Jose, CA, USA, May 1999.
- [200] C. Jörgensen, A. Jaimes, A. B. Benitez, and S.-F. Chang. A Conceptual Framework and Empirical Research for Classifying Visual Descriptors. *Journal of the American Society for Information Science and Technology*, 52(11):938–947, 2001.
- [201] C. Jörgensen and P. Jörgensen. Testing a vocabulary for image indexing and ground truthing. In G. Beretta and R. Schettini, editors, *Internet Imaging III*, volume 4672 of *SPIE Proceedings*, pages 207–215, San Jose, CA, USA, January 21–22 2002.
- [202] N. Kando. Overview of the Second NTCIR Workshop. In *Proceedings of the Second NTCIR Workshop on Research in Chinese & Japanese Text Retrieval and Text Summarization*, National Institute of Informatics, Tokyo, Japan, March 2001.
- [203] L. M. Kaplan, R. Murenzi, and K. R. Namuduri. Fast Texture Database Retrieval Using Extended Fractal Features. In I. K. Sethi and R. C. Jain, editors, *Storage and Retrieval for Image and Video Databases VI*, volume 3312 of *SPIE Proceedings*, pages 162–175, San Jose, CA, USA, January 28–30 1998.
- [204] D. Kapur, Y. N. Lakshman, and T. Saxena. Computing Invariants Using Elimination Methods. In *Proceedings of the International Symposium on Computer Vision (ISCV'95)*, pages 97–102, Coral Gables, FL, USA, November 21–23 1995.
- [205] T. Kato. Database Architecture for Content-Based Image Retrieval. In A. A. Jamberdino and W. Niblack, editors, *Image Storage and Retrieval Systems*,

volume 1662 of *SPIE Proceedings*, pages 112–123, San Jose, CA, USA, April 1992.

- [206] H. Kauppinen, T. Seppänen, and M. Pietikäinen. An Experimental Comparison of Autoregressive and Fourier-Based Descriptors in 2-D Shape Classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(2):201–207, 1995.
- [207] L. H. Keister. User Types and Queries: Impact on Image Access Systems. In R. Fidel, T. Bellardo Hahn, E. M. Rasmussen, and P. J. Smith, editors, *Challenges in Indexing Electronic Text and Images*, ASIS Monograph Series, pages 7–22. Learned Information, Inc., Medford, NJ, USA, 1994.
- [208] Y.-S. Kim and W.-Y. Kim. Content-Based Trademark Retrieval System Using Visually Salient Features. In *Proceedings of the 1997 IEEE Conference on Computer Vision and Pattern Recognition (CVPR '97)*, pages 307–312, San Juan, Puerto Rico, June 17–19 1997. IEEE Computer Society.
- [209] K. Kishida, K.-h. Chen, S. Lee, K. Kuriyama, N. Kando, H.-H. Chen, S. H. Myaeng, and K. Eguchi. Overview of CLIR Task at the Fourth NTCIR Workshop. In *Proceedings of the Fourth NTCIR Workshop on Research in Chinese & Japanese Text Retrieval and Text Summarization*, National Institute of Informatics, Tokyo, Japan, April 2003.
- [210] B. Ko, J. Peng, and H. Byun. A new content-based image retrieval system using hand gesture and relevance feedback. In *Proceedings of the Second International Conference on Multimedia and Exposition (ICME'2001)*, pages 501–504, Tokyo, Japan, August 22–25 2001. IEEE Computer Society.
- [211] S. Kopf, T. Haenselmann, and W. Effelsberg. Enhancing Curvature Scale Space Features for Robust Shape Classification. In *Proceedings of the 2005 International Conference on Multimedia and Exposition (ICME'2005)*, Am-

- sterdam, The Netherlands, July 6–8 2005. IEEE Computer Society. (CD-ROM).
- [212] M. Koskela, J. Laaksonen, S. Laakso, and E. Oja. Evaluating the performance of Content-Based Image Retrieval Systems. In R. Laurini, editor, *Fourth International Conference On Visual Information Systems (VISUAL'2000)*, volume 1929 of *Lecture Notes in Computer Science (LNCS)*, pages 430–441, Lyon, France, November 2–4 2000. Springer.
- [213] D. Koubaroulis, J. Matas, and J. Kittler. Evaluating colour-based object recognition algorithms on the soil-47 database. In D. Suter and A. Bab-Hadiashar, editors, *The Fifth Asian Conference on Computer Vision (ACCV 2002)*, pages 840–845, Melbourne, Australia, January 23–25 2002. Asian Federation of Computer Vision Societies.
- [214] W. Kraaij, A. F. Smeaton, and P. Over. TRECVID 2004 - An Overview. In *Online Proceedings of the TREC Video Retrieval Evaluation 2004*, Gaithersburg, MD, USA, November 15–16 2004.
- [215] W. Kraaij, A. F. Smeaton, and P. Over. TRECVID 2006 - An Introduction. In *Online Proceedings of the TREC Video Retrieval Evaluation 2006*, Gaithersburg, MD, USA, November 13–14 2006.
- [216] S. Krishnamachari and R. Chellappa. Multiresolution Gauss-Markov Random Field Models for Texture Segmentation. *IEEE Transactions on Image Processing*, 6(2):251–267, February 1997.
- [217] K. Kuriyama, N. Kando, T. Nozue, and K. Eguchi. Pooling for a Large-Scale Test Collection: An Analysis of the Search Results from the First NTCIR Workshop. *Information Retrieval*, 5(1):41–59, 2002.
- [218] R. Kurniawati, J. S. Jin, and J. Shepherd. SS+-Tree: An Improved Index Structure for Similarity Searches in a High-Dimensional Feature Space. In

- I. K. Sethi and R. C. Jain, editors, *Storage and Retrieval for Image and Video Databases V*, volume 3022 of *SPIE Proceedings*, pages 110–120, San Jose, CA, USA, February 1997.
- [219] J. Kustanowitz and B. Shneiderman. Motivating Annotation for Personal Digital Photo Libraries: Lowering Barriers While Raising Incentives. Technical Report HCIL-2004-18, University of Maryland, College Park, MD, USA, January 2005.
- [220] T. Kuyel and J. Ghosh. A fast space localized computation of the outputs of a Gabor filter bank. In *Proceedings of the IASTED Conference on Signal and Image Processing (SIP'95)*, pages 511–514, Las Vegas, NV, USA, November 20–23 1995.
- [221] K.-L. Kwok, L. Grunfeld, N. Dinstl, and P. Deng. TREC2005 Robust Track Experiments using PIRCS. In E. M. Voorhees and L. P. Buckland, editors, *Proceedings of the Fourteenth Text REtrieval Conference (TREC 2005)*, Gaithersburg, MD, USA, November 15–18 2005.
- [222] A. Laine and J. Fan. Texture Classification by Wavelet Packet Signatures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(11):1186–1191, November 1993.
- [223] P. M. Lam, J. K. Wu, and B. M. Mehtre. STAR – A System for Trademark Archival and Retrieval. In *Proceedings of the Second Asian Conference on Computer Vision (ACCV 1995)*, pages 214–217, Singapore, December 5–8 1995.
- [224] B. Larsen, A. Trotman, B. Sigurbjörnsson, S. Geva, M. Lalmas, and S. Malik. INEX 2006 Guidelines for Topic Development. In N. Fuhr, M. Lalmas, and A. Trotman, editors, *INEX 2006 Workshop Pre-Proceedings*, pages 373–380, Schloss Dagstuhl, Germany, December 18–20 2006. DELOS: Network of Excellence on Digital Libraries.

- [225] R. R. Larson. Text Retrieval and Blind Feedback for the ImageCLEF Photo Task. In C. Peters, P. D. Clough, F. C. Gey, J. Karlgren, B. Magnini, D. W. Oard, M. de Rijke, and M. Stempfhuber, editors, *Evaluation of Multilingual and Multi-modal Information Retrieval: Seventh Workshop of the Cross-Language Evaluation Forum (CLEF 2006)*, Lecture Notes in Computer Science (LNCS), Alicante, Spain, September 19–21 2006. Springer. (in press).
- [226] V. Lavrenko, R. Manmatha, and J. Jeon. A Model for Learning the Semantics of Pictures. In S. Thrun, L. K. Saul, and B. Schölkopf, editors, *The Proceedings of the 2003 Conference on Advances in Neural Information Processing Systems (NIPS 16)*, pages 553–560, Vancouver, BC, Canada, December 8–13 2004. MIT Press.
- [227] S.-Y. Lee and F.-J. Hsu. 2D C-String: A New Spatial Knowledge Representation for Image Database Systems. *Pattern Recognition*, 23(10):1077–1087, 1990.
- [228] S.-Y. Lee, M. C. Yang, and J. W. Chen. 2D B-string: A Spatial Knowledge Representation for Image Database Systems. In *Proceedings of the Second International Computer Science Conference (ICSC'92)*, pages 609–615, Hong Kong, 1992.
- [229] T. M. Lehmann, T. Deselaers, H. Schubert, M. O. Güld, C. Thies, B. Fischer, and K. Spitzer. The IRMA Code for Unique Classification of Medical Images. In H. K. Huang and O. M. Ratib, editors, *Medical Imaging 2003: PACS and Integrated Medical Information Systems: Design and Evaluation*, volume 5033 of *SPIE Proceedings*, pages 440–451, San Diego, CA, USA, February 18 2003.
- [230] T. M. Lehmann, T. Deselaers, H. Schubert, M. O. Güld, C. Thies, B. Fischer, and K. Spitzer. IRMA - a content-based approach to Image Retrieval in Medical Applications. In *IRMA International Conference 2006*, pages 911–912, Washington, DC, USA, May 21–24 2006.

- [231] Z. Lei, D. Keren, and D. Cooper. Computationally Fast Bayesian Recognition of Complex Objects Based on Mutual Algebraic Invariants. In *Proceedings of the 1995 International Conference on Image Processing (ICIP'95)*, pages 635–638, Washington, DC, USA, October 23–26 1995.
- [232] C. H. C. Leung and D. Hibler. Architecture of a Pictorial Database Management System. Technical report, British Library Research, London, UK, 1991.
- [233] C. H. C. Leung, J. Hibler, and N. Mwaru. Content-based retrieval in multimedia databases. *SIGGRAPH Computer Graphics*, 28(1):24–28, 1994.
- [234] C. H. C. Leung and H. Ip. Benchmarking for Content-Based Visual Information Search. In R. Laurini, editor, *Fourth International Conference On Visual Information Systems (VISUAL'2000)*, volume 1929 of *Lecture Notes in Computer Science (LNCS)*, pages 442–456, Lyon, France, November 2–4 2000. Springer.
- [235] C. H. C. Leung and Z. J. Zheng. Image Data Modeling For Efficient Content Indexing. In *Proceedings of the International Workshop on Multi-Media Database Management Systems*, pages 143–150, Blue Mountain Lake, NY, USA, August 28–30 1995.
- [236] M. S. Lew. Next-Generation Web Searches for Visual Content. *IEEE Computer*, 33(11):46–53, November 2000.
- [237] M. S. Lew, N. Sebe, C. Djeraba, and R. C. Jain. Content-Based Multimedia Information Retrieval: State of the Art and Challenges. *ACM Transactions Multimedia Computing, Communications and Applications*, 2(1):1–19, February 2006.
- [238] C.-S. Li and V. Castelli. Deriving Texture Feature Set for Content-based Retrieval of Satellite Image Database. In *Proceedings of the 1997 IEEE Inter-*

national Conference on Image Processing (ICIP'97), pages 576–579, Washington, DC, USA, October 26–29 1997.

- [239] J. Li. A Mutual Semantic Endorsement Approach to Image Retrieval and Context Provision. In *Proceedings of the Seventh ACM SIGMM International Workshop on Multimedia Information Retrieval (MIR'05)*, pages 173–182, Singapore, November 10–11 2005. ACM Press.
- [240] J. Li and H.-H. Sun. On Interactive Browsing of Large Images. *IEEE Transactions on Multimedia*, 5(4):581–590, December 2003.
- [241] R. Li, B. Bhanu, and A. Dong. Coevolutionary Feature Synthesized EM Algorithm for Image Retrieval. In *Proceedings of the 13th ACM International Conference on Multimedia (ACM Multimedia 2005)*, pages 696–705, Singapore, November 6–11 2005. ACM Press.
- [242] H.-C. Lin, L.-L. Wang, and S.-N. Yang. Color Image Retrieval Based on Hidden Markov Models. In *Proceedings of the 1995 International Conference on Image Processing (ICIP'95)*, pages 342–345, Washington, DC, USA, October 23–26 1995.
- [243] K.-I. Lin, H. V. Jagadish, and C. Faloutsos. The TV-Tree: An Index Structure for High-Dimensional Data. *VLDB Journal: Very Large Data Bases*, 3(4):517–542, 1994.
- [244] Y.-Y. Lin, T.-L. Liu, and H.-T. Chen. Semantic Manifold Learning for Image Retrieval. In *Proceedings of the 13th ACM International Conference on Multimedia (ACM Multimedia 2005)*, pages 249–258, Singapore, November 6–11 2005. ACM Press.
- [245] F. Liu and R. W. Picard. Periodicity, Directionality, and Randomness: Wold Features for Image Modeling and Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(7):722–733, July 1996.

- [246] H. Liu, X. Xie, X. Tang, Z.-W. Li, and W.-Y. Ma. Effective Browsing of Web Image Search Results. In *Proceedings of the 6th ACM SIGMM International Workshop on Multimedia Information Retrieval (MIR'04)*, pages 84–90, New York, NY, USA, October 15–16 2004. ACM Press.
- [247] Y. Liu, D. Zhang, G. Lu, and W.-Y. Ma. Region-Based Image Retrieval with High-Level Semantic Color Names. In *Proceedings of the 11th International Multimedia Modelling Conference (MMM'05)*, pages 180–187, Melbourne, Australia, January 12–14 2005. IEEE Computer Society.
- [248] S. Loncaric. A Survey of Shape Analysis Techniques. *Pattern Recognition*, 31(8):983–1001, 1998.
- [249] F. Long, H. Zhang, and D. D. Feng. Fundamentals of Content-Based Image Retrieval. In D. D. Feng, W.-C. Siu, and H. Zhang, editors, *Multimedia Information Retrieval and Management: Technological Fundamentals and Applications*, pages 1–26. Springer, 2003.
- [250] J. B. Lovins. Development of a Stemming Algorithm. *Mechanical Translation and Computational Linguistics*, 11:22–31, 1968.
- [251] G. Lu and A. Sajjanhar. Region-based Shape Representation and Similarity Measure Suitable for Content-based Image Retrieval. *Multimedia Systems*, 7(2):165–174, 1999.
- [252] Y. Lu, C. Hu, X. Zhu, H. Zhang, and Q. Yang. A Unified Framework for Semantics and Feature Based Relevance Feedback in Image Retrieval Systems. In *Proceedings of the Eighth ACM International Conference on Multimedia (ACM Multimedia 2000)*, pages 31–37, Marina del Rey, CA, USA, December 1–6 2000. ACM Press.
- [253] C. Luoni. Development of an Interface to a Database Storing the Features of a Multimedia Retrieval System. License thesis (BSc), Computer Vision and

Multimedia Laboratory at the University of Geneva, Geneva, Switzerland, 2000.

- [254] W. Ma and B. Manjunath. Texture Features and Learning Similarity. In *Proceedings of the 1996 IEEE Conference on Computer Vision and Pattern Recognition (CVPR '96)*, pages 425–430, San Francisco, CA, USA, June 18–20 1996. IEEE Computer Society.
- [255] W.-Y. Ma and B. Manjunath. A Texture Thesaurus for Browsing Large Aerial Photographs. *Journal of the American Society of Information Science*, 49(7):633–648, 1998.
- [256] W.-Y. Ma and B. S. Manjunath. Edge Flow: A Framework of Boundary Detection and Image Segmentation. In *Proceedings of the 1997 IEEE Conference on Computer Vision and Pattern Recognition (CVPR '97)*, pages 744–749, San Juan, Puerto Rico, June 17–19 1997. IEEE Computer Society.
- [257] W.-Y. M. Ma and B. S. Manjunath. NeTra: A Toolbox for Navigating Large Image Databases. *Multimedia Systems*, 7(3):184–198, 1999.
- [258] N. Maillot, J.-P. Chevallet, V. Valea, and J. H. Lim. IPAL Inter-Media Pseudo-Relevance Feedback Approach to ImageCLEF 2006 Photo Retrieval. In *CLEF working notes*, Alicante, Spain, September 2006.
- [259] S. G. Mallat. A Theory for Multiresolution Signal Decomposition: The Wavelet Representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):674–693, July 1989.
- [260] B. B. Mandelbrot. *The Fractal Geometry of Nature*. Freeman and Co, New York, USA, 1983.
- [261] B. S. Manjunath, J.-R. Ohm, V. V. Vasudevan, and A. Yamada. Color and Texture Descriptors. *IEEE Transactions on Circuits and Systems for Video Technology*, 11(6):703–715, June 2001.

- [262] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, England, 2007.
- [263] J. Mao and A. K. Jain. Texture Classification and Segmentation Using Multiresolution Simultaneous Autoregressive Models. *Pattern Recognition*, 25(2):173–188, 1992.
- [264] K. Markey. Interindexer Consistency Tests: a literature review and report of a test of consistency in indexing visual materials. *Library and Information Science Research*, 6:155–177, 1984.
- [265] M. Markkula and E. Sormunen. Searching for Photos – Journalists’ Practices in Pictorial IR. In J. P. Eakins, D. J. Harper, and J. Jose, editors, *The Challenge of Image Retrieval, A Workshop and Symposium on Image Retrieval*, Electronic Workshops in Computing, pages 1–13, Newcastle upon Tyne, UK, February 5–6 1998. The British Computer Society.
- [266] M. Markkula and E. Sormunen. End-User Searching Challenges Indexing Practices in the Digital Newspaper Photo Archive. *Information Retrieval*, 1(4):259–285, 2000.
- [267] M. Markkula, M. Tico, B. Sepponen, K. Nirkkonen, and E. Sormunen. A Test Collection for the Evaluation of Content-Based Image Retrieval Algorithms - A User and Task-Based Approach. *Information Retrieval*, 4(3-4):275–293, 2001.
- [268] J. M. Martínez, R. Koenen, and F. Pereira. MPEG-7: The Generic Multimedia Content Description Standard. *IEEE Multimedia*, 9(2):78–87, 2002.
- [269] J. L. Martínez-Fernández, J. Villena, A. García-Serrano, and P. Martínez. Expanding Queries through Word Sense Disambiguation. In C. Peters, P. D. Clough, F. C. Gey, J. Karlgren, B. Magnini, D. W. Oard, M. de Rijke, and M. Stempfhuber, editors, *Evaluation of Multilingual and Multi-modal Infor-*

mation Retrieval: Seventh Workshop of the Cross-Language Evaluation Forum (CLEF 2006), Lecture Notes in Computer Science (LNCS), Alicante, Spain, September 19–21 2006. Springer. (in press).

- [270] E. Mathias and A. Conci. Comparing the Influence of Color Spaces and Metrics in Content-Based Image Retrieval. In *Proceedings of the International Symposium on Computer Graphics, Image Processing, and Vision (SIBGRAPI '98)*, pages 371–378, Rio de Janeiro, Brazil, October 20–23 1998.
- [271] J. Mauro. Oracle interMedia: Managing Multimedia Content. An Oracle White Paper, March 2001.
- [272] H. McCorry and I. O. Morrison. Report on the Catechism Project. Technical report, National Museums of Scotland, Chambers Street, Edinburgh, Scotland, 1995.
- [273] B. M. Mehtre, M. S. Kankanhalli, and W. F. Lee. Shape Measures for Content Based Image Retrieval: A Comparison. *Information Processing and Management*, 33(3):319–337, 1997.
- [274] D. Metzler and R. Manmatha. An Inference Network Approach to Image Retrieval. In *Proceedings of the Third International Conference on Image and Video Retrieval (CIVR 2004)*, volume 3115 of *Lecture Notes in Computer Science (LNCS)*, pages 42–50, Dublin, Ireland, July 21–23 2004. Springer.
- [275] K. Mikolajczyk and C. Schmid. A Performance Evaluation of Local Descriptors. In *Proceedings of the 2003 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2003)*, volume 2, pages 257–263, Madison, WI, USA, June 16–22 2003. IEEE Computer Society.
- [276] K. Mikolajczyk and C. Schmid. Scale & Affine Invariant Interest Point Detectors. *International Journal on Computer Vision*, 60(1):63–86, 2004.

- [277] S. Mizzaro. Relevance: The Whole History. *Journal of the American Society for Information Science*, 48(9):810–832, 1997.
- [278] S. Mizzaro. How many relevances in information retrieval? *Interacting with Computers*, 10(3):303–320, 1998.
- [279] P.-A. Moëllic and C. Fluhr. ImagEVAL 2006 Official campaign. Technical report, CEA List, Fontenay-Aux-Roses, France, December 19 2006.
- [280] A. Mojsilović, J. Kovačević, J. Hu, R. J. Safranek, and S. K. Ganapathy. Matching and Retrieval Based on the Vocabulary and Grammar of Color Patterns. *IEEE Transactions on Image Processing*, 9(1):38–54, January 2000.
- [281] A. Mojsilović and B. Rogowitz. Capturing Image Semantics with Low Level Descriptors. In *Proceedings of the 2001 IEEE International Conference on Image Processing (ICIP 2001)*, pages 18–21, Thessaloniki, Greece, October 7–10 2001. IEEE Computer Society.
- [282] F. Mokhtarian, S. Abbasi, and J. Kittler. Efficient and Robust Retrieval by Shape Content through Curvature Scale Space. In A. W. M. Smeulders and R. Jain, editors, *Proceedings of the International Workshop on Image Databases and Multi-Media Search*, pages 35–42, Amsterdam, The Netherlands, August 22–23 1996. Amsterdam University Press.
- [283] F. Monay and D. Gatica-Perez. On Image Auto-annotation With Latent Space Models. In *Proceedings of the Eleventh ACM International Conference on Multimedia (ACM Multimedia 2003)*, pages 275–278, Berkeley, CA, USA, November 2–8 2003. ACM Press.
- [284] G. Mori, S. Belongie, and J. Malik. Efficient Shape Matching Using Shape Contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(11):1832–1837, December 2005.

- [285] Y. Mori, H. Takahashi, and R. Oka. Image-to-word Transformation Based on Dividing and Vector Quantizing Images With Words. In *First International Workshop on Multimedia Intelligent Storage and Retrieval Management (MISRM'99)*, Orlando, FL, USA, October 30 1999.
- [286] J. Mothe and L. Tanguy. Linguistic features to predict query difficulty - a case study on previous TREC campaigns. In *ACM SIGIR Workshop: Predicting Query Difficulty - Methods and Applications*, pages 7–10, Salvador, Brazil, August 19 2005.
- [287] H. Müller. *User Interaction and Evaluation in Content-Based Visual Information Retrieval*. PhD thesis, Computer Vision and Multimedia Laboratory, University of Geneva, Geneva, Switzerland, June 2002.
- [288] H. Müller, P. D. Clough, A. Geissbuhler, and W. Hersh. ImageCLEF 2004-2005: results, experiences and new ideas for image retrieval evaluation. In *Proceedings of the Fourth International Workshop on Content-Based Multimedia Indexing (CBMI 2005)*. CD-ROM, Riga, Latvia, June 21–23 2005.
- [289] H. Müller, P. D. Clough, W. Hersh, T. Deselaers, T. Lehmann, and A. Geissbuhler. Using heterogeneous annotation and visual information for the benchmarking of image retrieval systems. In S. Santini, R. Schettini, and T. Gevers, editors, *Internet Imaging VII*, volume 6061 of *SPIE Proceedings*, page 606105, San Jose, CA, USA, January 16 2006.
- [290] H. Müller, T. Deselaers, T. Lehmann, P. D. Clough, E. Kim, and W. Hersh. Overview of the ImageCLEFmed 2006 Medical Retrieval and Medical Annotation Tasks. In C. Peters, P. D. Clough, F. C. Gey, J. Karlgren, B. Magnini, D. W. Oard, M. de Rijke, and M. Stempfhuber, editors, *Evaluation of Multilingual and Multi-modal Information Retrieval: Seventh Workshop of the Cross-Language Evaluation Forum (CLEF 2006)*, Lecture Notes in Computer Science (LNCS), Alicante, Spain, September 19–21 2006. Springer. (in press).

- [291] H. Müller and A. Geissbuhler. How to Visually Retrieve Images from the St. Andrews Collection Using GIFT. In C. Peters, P. D. Clough, J. Gonzalo, G. J. F. Jones, M. Kluck, and B. Magnini, editors, *Multilingual Information Access for Text, Speech and Images: Fifth Workshop of the Cross-Language Evaluation Forum (CLEF 2004)*, volume 3491 of *Lecture Notes in Computer Science (LNCS)*, pages 633–642, Bath, UK, September 15–17 2004. Springer.
- [292] H. Müller, A. Geissbuhler, and S. Marchand-Maillet. Extensions to the Multimedia Retrieval Markup Language – A Communication Protocol for Content-Based Image Retrieval. In *Proceedings of the International Workshop on Content-Based Multimedia Indexing (CBMI 2003)*, pages 173–180, Rennes, France, September 22–24 2003.
- [293] H. Müller, A. Geissbuhler, S. Marchand-Maillet, and P. D. Clough. Benchmarking Image Retrieval Applications. In J. K. Wu and C. H. C. Leung, editors, *The Seventh International Conference on Visual Information Systems (VIS'2004)*, pages 334–337, San Francisco, CA, USA, September 2004. Knowledge Systems Institute.
- [294] H. Müller, S. Marchand-Maillet, and T. Pun. The Truth About Corel – Evaluation in Image Retrieval. In *Proceedings of the International Conference on the Challenge of Image and Video Retrieval (CIVR 2002)*, volume 2383 of *Lecture Notes in Computer Science (LNCS)*, pages 38–49, London, UK, July 2002. Springer.
- [295] H. Müller, N. Michoux, D. Bandon, and A. Geissbuhler. A Review of Content-Based Image Retrieval Systems in Medicine – Clinical Benefits and Future Directions. *International Journal of Medical Informatics*, 73(1):1–23, 2004.
- [296] H. Müller, W. Müller, S. Marchand-Maillet, D. M. Squire, and T. Pun. Long Term Learning in Content-Based Image Retrieval. Technical Report 00.04, Computer Vision Group, Computing Centre, University of Geneva, rue Général Dufour, 24, CH-1211 Genève, Switzerland, February 2000.

- [297] H. Müller, W. Müller, S. Marchand-Maillet, D. M. Squire, and T. Pun. A web-based evaluation system for content-based image retrieval. In *Proceedings of the 9th ACM International Conference on Multimedia (ACM Multimedia 2001)*, pages 50–54, Ottawa, Canada, September 30 - October 5 2001. ACM Press.
- [298] H. Müller, W. Müller, S. Marchand-Maillet, D. M. Squire, and T. Pun. Automated Benchmarking in Content-Based Image Retrieval. In *Proceedings of the Second International Conference on Multimedia and Exposition (ICME'2001)*, pages 321–324, Tokyo, Japan, August 22–25 2001. IEEE Computer Society.
- [299] H. Müller, W. Müller, S. Marchand-Maillet, D. M. Squire, and T. Pun. A Framework for Benchmarking in Visual Information Retrieval. *International Journal on Multimedia Tools and Applications*, 21:55–73, 2003. (Special Issue on Multimedia Information Retrieval).
- [300] H. Müller, W. Müller, D. M. Squire, S. Marchand-Maillet, and T. Pun. Learning Feature Weights From User Behavior in Content-Based Image Retrieval. In S. J. Simoff and O. R. Zaiane, editors, *International Conference on Knowledge Discovery and Data Mining (Workshop on Multimedia Data Mining MDM/KDD2000)*, pages 67–72, Boston, MA, USA, August 20–23 2000.
- [301] H. Müller, W. Müller, D. M. Squire, S. Marchand-Maillet, and T. Pun. Strategies for Positive and Negative Relevance Feedback in Image Retrieval. In A. Sanfeliu, J. J. Villanueva, M. Vanrell, R. Alc  zar, J.-O. Eklundh, and Y. Aloimonos, editors, *Proceedings of the 15th International Conference on Pattern Recognition (ICPR'00)*, pages 1043–1046, Barcelona, Spain, September 3–8 2000. IEEE Computer Society.
- [302] H. Müller, W. Müller, D. M. Squire, S. Marchand-Maillet, and T. Pun. Performance Evaluation in Content-Based Image Retrieval: Overview and Proposals. *Pattern Recognition Letters*, 22(5):593–601, April 2001.

- [303] H. Müller, W. Müller, D. M. Squire, Z. Pečenović, S. Marchand-Maillet, and T. Pun. An Open Framework for Distributed Multimedia Retrieval. In *Recherche d'Informations Assistée par Ordinateur (RIAO'2000) Computer-Assisted Information Retrieval*, volume 1, pages 701–712, Paris, France, April 12–14 2000.
- [304] H. Müller, A. Rosset, J.-P. Vallée, F. Terrier, and A. Geissbuhler. A reference data set for the evaluation of medical image retrieval systems. *Journal of Computerized Medical Imaging and Graphics*, 28:65–77, 2004.
- [305] H. Müller, D. M. Squire, W. Müller, and T. Pun. Efficient Access Methods for Content-Based Image Retrieval With Inverted Files. In S. Panchanathan, S.-F. Chang, and C.-C. J. Kuo, editors, *Multimedia Storage and Archiving Systems IV (VV02)*, volume 3846 of *SPIE Proceedings*, pages 461–472, Boston, MA, USA, September 20–22 1999.
- [306] H. Müller, D. M. Squire, and T. Pun. Learning from User Behavior in Image Retrieval: Application of the Market Basket Analysis. *International Journal of Computer Vision*, 56(1–2):65–77, 2004. (Special Issue on Content-Based Image Retrieval).
- [307] W. Müller. *Design and Implementation of a Flexible Content-Based Image Retrieval Framework - The GNU Image Finding Tool*. PhD thesis, Computer Vision and Multimedia Laboratory, University of Geneva, Geneva, Switzerland, September 2001.
- [308] W. Müller, S. Marchand-Maillet, H. Müller, and T. Pun. Towards a fair benchmark for image browsers. In J. R. Smith, C. Le, S. Panchanathan, and C.-C. J. Kuo, editors, *Internet Multimedia Management Systems*, volume 4210 of *SPIE Proceedings*, pages 262–271, Boston, MA, USA, October 2000.
- [309] W. Müller, H. Müller, S. Marchand-Maillet, T. Pun, D. M. Squire, Z. Pečenović, C. Giess, and A. P. de Vries. MRML: A Communication Proto-

- col for Content-Based Image Retrieval. In R. Laurini, editor, *Fourth International Conference On Visual Information Systems (VISUAL'2000)*, volume 1929 of *Lecture Notes in Computer Science (LNCS)*, pages 300–311, Lyon, France, November 2–4 2000. Springer.
- [310] W. Müller, Z. Pečenović, H. Müller, S. Marchand-Maillet, T. Pun, D. M. Squire, A. P. D. Vries, and C. Giess. MRML: An Extensible Communication Protocol for Interoperability and Benchmarking of Multimedia Information Retrieval Systems. In J. R. Smith, C. Le, S. Panchanathan, and C.-C. J. Kuo, editors, *Internet Multimedia Management Systems*, volume 4210 of *SPIE Proceedings*, pages 124–133, Boston, MA, USA, October 2000.
- [311] M. Nakazato and T. S. Huang. 3D MARS: Immersive Virtual Reality for Content-Based Image Retrieval. In *Proceedings of the Second International Conference on Multimedia and Exposition (ICME'2001)*, pages 45–48, Tokyo, Japan, August 22–25 2001. IEEE Computer Society.
- [312] A. D. Narasimhalu, M. S. Kankanhalli, and J.-K. Wu. Benchmarking Multimedia Databases. *Multimedia Tools and Applications*, 4(3):333–356, 1997.
- [313] S. A. Nene, S. K. Nayar, and H. Murase. Columbia Object Image Library (COIL-100). Technical Report Number CUCS-006-96, Department of Computer Science, Columbia University, New York, NY, USA, 1996.
- [314] S. A. Nene, S. K. Nayar, and H. Murase. Columbia Object Image Library (COIL-20). Technical Report Number CUCS-005-96, Department of Computer Science, Columbia University, New York, NY, USA, 1996.
- [315] N. Nes and M. Kersten. The Acoi Algebra: a Query Algebra for Image Retrieval Systems. In *Proceedings of the British National Conference on Databases (BNCOD)*, volume 1405 of *Lecture Notes in Computer Science/Lecture Notes in Artificial Intelligence (LNCS/LNAI)*, pages 77–88, Cardiff, Wales, UK, July 6–8 1998. Springer.

- [316] D. Neumann and K. R. Gegenfurtner. Image Retrieval and Perceptual Similarity. *ACM Transactions on Applied Perception (TAP)*, 3(1):31–47, January 2006.
- [317] R. T. Ng and A. Sedighian. Evaluating Multidimensional Indexing Structures for Images Transformed by Principal Component Analysis. In I. K. Sethi and R. C. Jain, editors, *Storage and Retrieval for Image and Video Databases IV*, volume 2670 of *SPIE Proceedings*, pages 50–61, San Jose, CA, USA, March 1996.
- [318] W. Niblack, R. Barber, W. Equitz, M. D. Flickner, E. H. Glasman, D. Petkovic, P. Yanker, C. Faloutsos, and G. Taubin. QBIC project: querying images by content, using color, texture, and shape. In W. Niblack, editor, *Storage and Retrieval for Image and Video Databases*, volume 1908 of *SPIE Proceedings*, pages 173–187, San Jose, CA, USA, April 1993.
- [319] J. Nielsen. *Usability Engineering*. Morgan Kaufmann, San Francisco, CA, USA, 1994.
- [320] J. Nievergelt, H. Hinterberger, and K. C. Sevcik. The Grid File: An Adaptable, Symmetric Multikey File Structure. *ACM Transactions on Database Systems*, 9(1):38–71, March 1984.
- [321] P. Niyogi. The Informational Complexity of Learning from Examples. Technical report, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, USA, September 1996.
- [322] M. Nölle and A. Hanbury. MUSCLE Coin Images Seibersdorf (CIS) Benchmark Competition 2006. *IAPR Newsletter April 2006*, 28(2):18–19, 2006.
- [323] M. Nölle and M. Rubik. Results of the MUSCLE CIS Coin Competition 2006. In M. Nölle, M. Rubik, and A. Hanbury, editors, *Proceedings of the MUSCLE CIS Coin Competition Workshop*, pages 1–5, Berlin, Germany, September 11 2006. Seibersdorf Research.

- [324] B. O'Connor, M. O'Connor, and J. Abbas. User Reactions as Access Mechanism: An Exploration Based on Captions for Images. *Journal of the American Society for Information Science*, 50(8):681–697, June 1999.
- [325] S. Ornager. Image retrieval: Theoretical and Empirical User Studies on Accessing Information in Images. In *Proceedings of the 60th Annual Meeting of the American Society for Information Science (ASIS 97)*, volume 34, pages 202–211, Washington, DC, USA, November 1–6 1997.
- [326] M. Ortega, Y. Rui, K. Chakrabarti, K. Porkaew, S. Mehrotra, and T. S. Huang. Supporting Ranked Boolean Similarity Queries in MARS. *IEEE Transactions on Knowledge and Data Engineering*, 10(6):905–925, December 1998.
- [327] M. Ortega-Binderberger, S. Mehrotra, K. Chakrabarti, and K. Porkaew. Web-MARS: A Multimedia Search Engine. In G. Beretta and J. J. McCann, editors, *Internet Imaging*, volume 3963 of *SPIE Proceedings*, pages 216–224, San Jose, CA, USA, January 23–28 2000. (SPIE Photonics West Conference).
- [328] P. Over, T. Ianeva, W. Kraaij, and A. F. Smeaton. TRECVID 2005 - An Overview. In *Online Proceedings of the TREC Video Retrieval Evaluation 2005*, Gaithersburg, MD, USA, November 14–15 2005.
- [329] P. Over, C. H. C. Leung, H. Ip, and M. Grubinger. Multimedia Retrieval Benchmarks. *Digital Multimedia on Demand, IEEE Multimedia April - June*, pages 80–84, 2004.
- [330] B. Ozer, W. Wolf, and A. N. Akansu. A Graph Based Object Description for Information Retrieval in Digital Image and Video Libraries. In *IEEE Workshop on Content-based Access of Image and Video Libraries (CBAIVL'99)*, pages 79–83, Fort Collins, CO, USA, June 22 1999. IEEE Computer Society.
- [331] C. D. Paice. Another Stemmer. *SIGIR Forum*, 24(3):56–61, 1990.

- [332] E. Panofsky. *Meaning in the Visual Arts: Papers in and on Art History*. Doubleday Anchor Books, Garden City, NY, USA, 1955.
- [333] G. Pass and R. Zabih. Histogram Refinement for Content-Based Image Retrieval. In *Proceedings of the 3rd IEEE Workshop on Applications of Computer Vision (WACV '96)*, pages 96–102, Washington, DC, USA, December 2–4 1996. IEEE Computer Society.
- [334] G. Pass and R. Zabih. Comparing Images Using Joint Histograms. *Multimedia Systems*, 7(3):234–240, 1999.
- [335] A. Pentland. Fractal-based Description of Natural Scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:661–674, 1984.
- [336] A. Pentland, R. W. Picard, and S. Sclaroff. Photobook: Content-Based Manipulation of Image Databases. *International Journal of Computer Vision*, 18(3):233–254, June 1996.
- [337] E. Persoon and K. S. Fu. Shape Discrimination Using Fourier Descriptors. *IEEE Transactions on Systems, Man and Cybernetics*, 7:170–179, March 1977.
- [338] C. Peters and M. Braschler. Cross Language System Evaluation: The CLEF Campaigns. *Journal of the American Society for Information Science and Technology*, 22(12):1067–1072, 2001.
- [339] C. Peters, P. D. Clough, F. C. Gey, J. Karlgren, B. Magnini, D. W. Oard, M. de Rijke, and M. Stempfhuber, editors. *Evaluation of Multilingual and Multi-modal Information Retrieval: Seventh Workshop of the Cross-Language Evaluation Forum (CLEF 2006)*, Lecture Notes in Computer Science (LNCS), Alicante, Spain, September 19–21 2006. Springer.
- [340] C. Peters, P. D. Clough, J. Gonzalo, G. J. F. Jones, M. Kluck, and B. Magnini, editors. *Multilingual Information Access for Text, Speech and Images: Fifth*

- Workshop of the Cross-Language Evaluation Forum (CLEF 2004)*, volume 3491 of *Lecture Notes in Computer Science (LNCS)*, Bath, UK, September 15–17 2004. Springer.
- [341] C. Peters, F. C. Gey, J. Gonzalo, H. Müller, G. J. F. Jones, M. Kluck, B. Magnini, M. de Rijke, and D. Giampiccolo, editors. *Accessing Multilingual Information Repositories: Sixth Workshop of the Cross-Language Evaluation Forum (CLEF 2005)*, volume 4022 of *Lecture Notes in Computer Science (LNCS)*, Vienna, Austria, September 20–22 2005. Springer.
- [342] C. Peters, J. Gonzalo, M. Braschler, and M. Kluck, editors. *Comparative Evaluation of Multilingual Information Access Systems: Fourth Workshop of the Cross-Language Evaluation Forum (CLEF 2003)*, volume 3237 of *Lecture Notes in Computer Science (LNCS)*, Trondheim, Norway, August 21–22 2003. Springer.
- [343] T. Pfund and S. Marchand-Maillet. Dynamic multimedia annotation tool. In G. Beretta and R. Schettini, editors, *Internet Imaging III*, volume 4672 of *SPIE Proceedings*, pages 216–224, San Jose, CA, USA, January 21–22 2002.
- [344] C. Picault. Constitution of the ImageEval Corpus : An end user-oriented approach. Technical report, Laboratoire Paragraphe – Université Paris 8, France, December 2006.
- [345] S. P. Pollard and A. W. Biermann. A Measure of Semantic Complexity for Natural Language Systems. In *Workshop on Syntactic and Semantic Complexity in Natural Language Processing Systems (NAACL-ANLP 2000)*, pages 42–46, Seattle, WA, USA, April 30 2000. Association for Computational Linguistics.
- [346] M. F. Porter. An Algorithm For Suffix Stripping. *Program*, 14(3):130–137, July 1980.

- [347] J. Preece. *Human-Computer Interaction*. Addison-Wesley, Harlow, England, 1994.
- [348] R. J. Prokop and A. P. Reeves. A Survey of Moment-Based Techniques for Unoccluded Object Representation and Recognition. *CVGIP: Graphical Models and Image Processing*, 54(5):438–460, September 1992.
- [349] F. Qian, M. Li, W.-Y. Ma, F. Ling, and B. Zhang. Alternating features spaces in relevance feedback. In *Proceedings of the ACM Multimedia Workshop on Multimedia Information Retrieval (MIR 2001)*, pages 14–17, Ottawa, Canada, October 5 2001. ACM Press.
- [350] M. M. Rahman, V. Sood, B. C. Desai, and P. Bhattacharya. CINDI at ImageCLEF 2006: Image Retrieval and Annotation Tasks for the General Photographic and Medical Image Collections. In C. Peters, P. D. Clough, F. C. Gey, J. Karlgren, B. Magnini, D. W. Oard, M. de Rijke, and M. Stempfhuber, editors, *Evaluation of Multilingual and Multi-modal Information Retrieval: Seventh Workshop of the Cross-Language Evaluation Forum (CLEF 2006)*, Lecture Notes in Computer Science (LNCS), Alicante, Spain, September 19–21 2006. Springer. (in press).
- [351] R. Rahmani, S. A. Goldman, H. Zhang, J. Krettek, and J. E. Fritts. Localized content based image retrieval. In *Proceedings of the Seventh ACM SIGMM International Workshop on Multimedia Information Retrieval (MIR'05)*, pages 227–236, Singapore, November 10–11 2005. ACM Press.
- [352] T. Randen and H. Husøy. Filtering for Texture Classification: A Comparative Study. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(4):291–310, April 1999.
- [353] E. M. Rasmussen. Indexing Images. *Annual Review of Information Science and Technology*, 32:169–196, 1997.

- [354] S. Ravela, R. Manmatha, and E. M. Riseman. Image Retrieval Using Scale Space Matching. In B. Buxton and R. Cippola, editors, *4th European Conference on Computer Vision (ECCV 1996)*, volume 1064 of *Lecture Notes in Computer Science (LNCS)*, pages 273–282, Cambridge, UK, April 15–18 1996. Springer.
- [355] N. H. Reid. The Photographic Collections in St Andrews University Library. *Scottish Archives*, 5:83–90, 1999.
- [356] E. S. Ristad. *The Language Complexity Game*. MIT Press, Cambridge, MA, USA, March 1993.
- [357] A. R. Robertson. Historical development of CIE Recommended Color Difference Equations. *COLOR research and applications*, 15(3):167–170, 1990.
- [358] S. E. Robertson and K. Spärck Jones. Simple, proven approaches to text retrieval. Technical Report 356, Computer Laboratory, University of Cambridge, 1994.
- [359] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. In D. K. Harman, editor, *Proceedings of the Third Text REtrieval Conference (TREC-3)*, pages 109–126, Gaithersburg, MD, USA, November 2–4 1994. Department of Commerce, National Institute of Standards and Technology.
- [360] J. T. Robinson. The K-D-B-Tree: A Search Structure for Large Multidimensional Dynamic Indexes. In *Proceedings of the 1981 ACM SIGMOD International Conference on Management of Data (SIGMOD '81)*, pages 10–18, Ann Arbor, MI, USA, April 29 – May 1 1981. ACM Press.
- [361] J. J. Rocchio. Relevance Feedback in Information Retrieval. In *The SMART Retrieval System, Experiments in Automatic Document Processing*, pages 313–323. Prentice Hall, Englewood Cliffs, NJ, USA, 1971.

- [362] K. Rodden, W. Basalaj, D. Sinclair, and K. R. Wood. Does Organisation by Similarity Assist Image Browsing? In *Proceedings of the SIGCHI Conference on Human Factors in Computing systems (CHI 2001)*, pages 190–197, Seattle, WA, USA, March 31 – April 5 2001. ACM Press.
- [363] K. Rodden and K. R. Wood. How Do People Manage Their Digital Photographs? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2003)*, pages 409–416, Fort Lauderdale, FL, USA, April 5–10 2003. ACM Press.
- [364] T. Rose, D. Elworthy, A. Kotcheff, A. Clare, and P. Tsonis. ANVIL: a system for the retrieval of captioned images using NLP techniques. In *Proceedings of the Challenge of Image Retrieval Conference 2000 (CIR'2000)*, Brighton, UK, May 4–5 2000.
- [365] A. Rosset, H. Müller, M. Martins, N. Dfouni, J.-P. Vallée, and O. Ratib. Casimage Project - A Digital Teaching Files Authoring Environment. *Journal of Thoracic Imaging*, 19(2):103–108, April 2004.
- [366] N. Roussopoulos, C. Faloutsos, and T. K. Sellis. An Efficient Pictorial Database System for PSQL. *IEEE Transactions on Software Engineering*, 14(5):639–650, May 1988.
- [367] Y. Rui and T. S. Huang. Optimizing Learning in Image Retrieval. In *Proceedings of the 2000 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2000)*, pages 236–245, Hilton Head Island, SC, USA, June 13–15 2000. IEEE Computer Society.
- [368] Y. Rui, T. S. Huang, and S.-F. Chang. Image Retrieval: Current Techniques, Promising Directions and Open Issues. *Journal of Visual Communication and Image Representation*, 10(4):39–62, April 1999.
- [369] Y. Rui, T. S. Huang, and S. Mehrotra. Content-based Image Retrieval With Relevance Feedback in MARS. In *Proceedings of the 1997 IEEE International*

Conference on Image Processing (ICIP'97), pages 815–818, Washington, DC, USA, October 26–29 1997.

- [370] Y. Rui, T. S. Huang, and S. Mehrotra. Relevance Feedback Techniques in Interactive Content–Based Image Retrieval. In I. K. Sethi and R. C. Jain, editors, *Storage and Retrieval for Image and Video Databases VI*, volume 3312 of *SPIE Proceedings*, pages 25–36, San Jose, CA, USA, January 28–30 1998.
- [371] Y. Rui, T. S. Huang, M. Ortega, and S. Mehrotra. Relevance Feedback: A Power Tool for Interactive Content–Based Image Retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 8(5):644–655, September 1998. (Special Issue on Segmentation, Description, and Retrieval of Video Content).
- [372] T. Sakai. Evaluating Evaluation Metrics based on the Bootstrap. In S. Dumais, E. N. Efthimiadis, D. Hawking, and K. Järvelin, editors, *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2006)*, pages 525–532, Seattle, WA, USA, August 6–11 2006. ACM Press.
- [373] T. Sakai. On the reliability of information retrieval metrics based on graded relevance. *Information Processing and Management*, 43:531–548, 2007.
- [374] P. S. Salembier, T. Sikora, and B. Manjunath, editors. *Introduction to MPEG-7: Multimedia Content Description Interface*. John Wiley & Sons, Inc., New York, NY, USA, 2002.
- [375] G. Salton and C. Buckley. Improving Retrieval Performance by Relevance Feedback. *Journal of the American Society for Information Science*, 41(4):288–287, 1990.
- [376] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill Book Company, New York City, NY, USA, 1983.

- [377] H. Samet. The Quadtree and Related Hierarchical Data Structures. *ACM Computing Surveys*, 16(2):187–260, 1984.
- [378] M. Sanderson and H. Joho. Forming Test Collections with No System Pooling. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2004)*, pages 33–40, Sheffield, UK, July 25–29 2004. ACM Press.
- [379] M. Sanderson and J. Kohler. Analyzing Geographic Queries. In *Online Proceedings of the Workshop on Geographic Information Retrieval at SIGIR 2004 (GIR '04)*, Sheffield, UK, July 25–29 2004.
- [380] M. Sanderson, J. Tian, and P. D. Clough. Testing an Automatic Organisation of Retrieved Images Into a Hierarchy. In *International Workshop On-toImage'2006 Language Resources for Content-Based Image Retrieval, held in conjunction with LREC'06*, pages 44–50, Genoa, Italy, May 22 2006.
- [381] M. Sanderson and J. Zobel. Information Retrieval System Evaluation: Effort, Sensitivity, and Reliability. In G. Marchionini, A. Moffat, J. Tait, R. Baeza-Yates, and N. Ziviani, editors, *Proceedings of the Twenty-Eighth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2005)*, pages 162–169, Salvador, Brazil, August 15–19 2005. ACM Press.
- [382] C. Santamaria, M. Bober, and W. Szajnowski. Texture Analysis using Level-crossing Statistics. In *Proceedings of the 17th International Conference on Pattern Recognition, Volume 2 (ICPR'04)*, pages 712–715, Cambridge, UK, August 23–26 2004. IEEE Computer Society.
- [383] S. Santini and R. C. Jain. Similarity Queries in Image Databases. In *Proceedings of the 1996 IEEE Conference on Computer Vision and Pattern Recognition (CVPR '96)*, pages 646–651, San Francisco, CA, USA, June 18–20 1996. IEEE Computer Society.

- [384] S. Santini and R. C. Jain. The Graphical Specification of Similarity Queries. *Journal of Visual Languages and Computing*, 7(4):403–421, 1996.
- [385] R. Schettini, G. Ciocca, and S. Zuffi. A Survey of Methods for Colour Image Indexing and Retrieval in Image Databases. In L. W. MacDonald and M. R. Luo, editors, *Color Imaging Science: Exploiting Digital Media*. John Wiley, 2001.
- [386] C. Schmid. A Structured Probabilistic Model for Recognition. In *Proceedings of the 1999 IEEE Conference on Computer Vision and Pattern Recognition (CVPR '99)*, pages 485–490, Fort Collins, CO, USA, June 23–25 1999. IEEE Computer Society.
- [387] m. c. schraefel, M. Karam, and S. Zhao. mSpace: interaction design for user-determined, adaptable domain exploration in hypermedia. In P. de Bra, editor, *Proceedings of AH 2003: Workshop on Adaptive Hypermedia and Adaptive Web Based Systems*, pages 217–235, Nottingham, UK, August 26–30 2003.
- [388] A. T. G. Schreiber, B. Dubbeldam, J. Wielemaker, and B. Wielinga. Ontology-Based Photo Annotation. *IEEE Intelligent Systems*, 16(3):66–74, 2001.
- [389] K. A. Schroeder. Layered indexing of images. *The Indexer*, 21(1):11–14, 1998.
- [390] S. Sclaroff, M. La Cascia, S. Sethi, and L. Taycher. Unifying Textual and Visual Cues for Content-Based Image Retrieval on the World Wide Web. *Computer Vision and Image Understanding*, 75(1/2):86–98, July/August 1999.
- [391] S. Sclaroff, L. Taycher, and M. La Cascia. ImageRover: a content-based Browser for the World Wide Web. In *IEEE Workshop on Content-based Access of Image and Video Libraries (CBAIVL'97)*, pages 2–9, San Juan, Puerto Rico, June 20 1997. IEEE Computer Society.

- [392] G. J. Scott and C.-R. Shyu. EBS k-d Tree: An Entropy Balanced Statistical k-d Tree for Image Databases with Ground-Truth Labels. In *Proceedings of the Second International Conference on Image and Video Retrieval (CIVR 2003)*, volume 2728 of *Lecture Notes in Computer Science (LNCS)*, pages 467–476, Urbana, IL, USA, July 24–25 2003. Springer.
- [393] T. B. Sebastian, P. N. Klein, and B. B. Kimia. On Aligning Curves. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(1):116–125, January 2003.
- [394] N. Sebe and M. S. Lew. Color-Based Retrieval. *Pattern Recognition Letters*, 22:223–230, 2001.
- [395] T. K. Sellis, N. Roussopoulos, and C. Faloutsos. The R+ Tree: A Dynamic Index for Multi-Dimensional Objects. In P. M. Stocker, W. Kent, and P. Hammersley, editors, *Proceedings of the 13th International Conference on Very Large Data Bases (VLDB'87)*, pages 507–518, Brighton, UK, September 1–4 1987. Morgan Kaufmann.
- [396] S. Shatford. Analyzing the Subject of a Picture: A Theoretical Approach. *Cataloging and Classification Quarterly*, 6(3):39–61, 1986.
- [397] J. Shi and J. Malik. Normalized Cuts and Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, August 2000.
- [398] C.-R. Shyu, C. E. Brodley, A. C. Kak, A. Kosaka, A. M. Aisen, and L. S. Broderick. ASSERT: A Physician-in-the-Loop Content-Based Retrieval System for HRCT Image Databases. *Computer Vision and Image Understanding*, 75(1–2):111–132, July/August 1999.
- [399] A. Singhal, C. Buckley, and M. Mitra. Pivoted Document Length Normalization. In *Proceedings of the 19th Annual International ACM SIGIR Conference*

on Research and Development in Information Retrieval (SIGIR '96), pages 21–29, Zurich, Switzerland, August 18–22 1996. ACM Press.

- [400] J. Sledge. Points of View. In D. Bearman, editor, *Multimedia Computing and Museums: Selected Papers from the Third International Conference on Hypermedia and Interactivity in Museums (ICHIM 95/MCN 95)*, pages 335–346, San Diego, CA, USA, October 9–13 1995. Archives & Museum Informatics.
- [401] A. F. Smeaton, W. Kraaij, and P. Over. TRECVID 2003 - An Overview. In *Online Proceedings of the TREC Video Retrieval Evaluation 2003*, Gaithersburg, MD, USA, November 17–18 2003.
- [402] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation Campaigns and TRECVID. In *Proceedings of the 8th ACM international workshop on Multimedia information retrieval (MIR '06)*, pages 321–330, Santa Barbara, CA, USA, 2006. ACM Press.
- [403] A. F. Smeaton, P. Over, and R. Taban. The TREC–2001 Video Track Report. In E. M. Voorhees, editor, *Proceedings of the Tenth Text REtrieval Conference (TREC 2001)*, pages 52–60, Gaithersburg, MD, USA, April 18 2002. Department of Commerce, National Institute of Standards and Technology.
- [404] A. F. Smeaton, P. Over, and R. Taban. The TREC–2002 Video Track Report. In E. M. Voorhees, editor, *Proceedings of the Eleventh Text REtrieval Conference (TREC 2002)*, pages 69–85, Gaithersburg, MD, USA, March 5 2003. Department of Commerce, National Institute of Standards and Technology.
- [405] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. C. Jain. Content-Based Image Retrieval at the End of the Early Years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, 2000.
- [406] J. R. Smith. *Integrated Spatial and Feature Image Systems: Retrieval, Analysis and Compression*. PhD thesis, Graduate School of Arts and Sciences, Columbia University, New York, NY, 1997.

- [407] J. R. Smith. Image Retrieval Evaluation. In *IEEE Workshop on Content-based Access of Image and Video Libraries (CBAIVL'98)*, pages 112–113, Santa Barbara, CA, USA, June 21 1998.
- [408] J. R. Smith and S.-F. Chang. Automated Binary Texture Feature Sets For Image Retrieval. In *Proceedings of the 1996 IEEE International Conference On Acoustics, Speech and Signal Processing*, pages 2239–2242, Atlanta, GA, USA, May 7–10 1996.
- [409] J. R. Smith and S.-F. Chang. Tools and Techniques for Color Image Retrieval. In I. K. Sethi and R. C. Jain, editors, *Storage and Retrieval for Image and Video Databases IV*, volume 2670 of *SPIE Proceedings*, pages 426–437, San Jose, CA, USA, March 1996.
- [410] J. R. Smith and S.-F. Chang. VisualSEEK: a fully automated content-based image query system. In *Proceedings of the fourth ACM international conference on Multimedia (MULTIMEDIA '96)*, pages 87–98, Boston, MA, USA, November 20 1996. ACM Press.
- [411] J. R. Smith and S.-F. Chang. Querying by Color Regions Using the VisualSEEK Content-based Visual Query System. In *Proceedings of the IJCAI Workshop on Intelligent Multimedia Information Retrieval*, pages 23–41, Nagoya, Japan, August 23–29 1997.
- [412] W. W. S. So, C. H. C. Leung, and Z. J. Zheng. Search Space Reduction Techniques for Image Databases. In *Proceedings of the First International Workshop on Image Databases and Multimedia Search (IDB-MMS'1996)*, pages 179–186, Amsterdam, The Netherlands, August 22–23 1996.
- [413] I. Soboroff, C. Nicholas, and P. Cahan. Ranking Retrieval Systems without Relevance Judgments. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*

(*SIGIR 2001*), pages 66–73, New Orleans, LA, USA, September 9–13 2001. ACM Press.

- [414] K. Spärck Jones. *Information Retrieval Experiment*. Butterworths, London, UK, 1981.
- [415] K. Spärck Jones and C. J. van Rijsbergen. Report on the need for and provision of an ideal information retrieval test collection. British Library Research and Development Report 5266, Computer Laboratory, University of Cambridge, 1975.
- [416] K. Spärck Jones and C. J. van Rijsbergen. Information Retrieval Test Collections. *Journal of Documentation*, 32:59–75, 1976.
- [417] K. Spärck Jones, S. Walker, and S. E. Robertson. A Probabilistic Model of Information Retrieval: Development and Comparative Experiments (Part 1). *Information Processing and Management*, 36:779–808, 2000.
- [418] K. Spärck Jones, S. Walker, and S. E. Robertson. A Probabilistic Model of Information Retrieval: Development and Comparative Experiments (Part 2). *Information Processing and Management*, 36:809–840, 2000.
- [419] A. Spink, H. Greisdorf, and J. Bateman. From highly relevant to not relevant: examining different regions of relevance. *Information Processing and Management*, 34(5):599–621, 1998.
- [420] A. Spink, B. J. Jansen, D. Wolfram, and T. Saracevic. From E-Sex to E-Commerce: Web Search Changes. *Computer*, 35(3):107–109, March 2002.
- [421] A. Spink, H. Partridge, and B. J. Jansen. Sexual and Pornographic Web Searching: Trends Analysis. *First Monday*, 11(9), September 2006.
- [422] A. Spink, D. Wolfram, B. J. Jansen, and T. Saracevic. Searching the Web: The Public And Their Queries. *Journal of the American Society for Information Science and Technology*, 52(3):226–234, February 2001.

- [423] A. Spink and J. L. Xu. Selected Results from a Large Study of Web Searching: The Excite Study. *Information Research*, 6(1):Digital Proceedings, October 2000.
- [424] D. M. Squire. General Methods For Invariant Pattern Recognition. In *Model-based Neural Networks for Invariant Pattern Recognition*, chapter 2, pages 8–33. School of Computing, Curtin University of Technology, Perth, Western Australia, October 1996.
- [425] D. M. Squire, H. Müller, and W. Müller. Improving Response Time by Search Pruning in a Content-Based Image Retrieval System, Using Inverted File Techniques. In *IEEE Workshop on Content-based Access of Image and Video Libraries (CBAIVL'99)*, pages 45–49, Fort Collins, CO, USA, June 22 1999. IEEE Computer Society.
- [426] D. M. Squire, W. Müller, and H. Müller. Relevance Feedback and Term Weighting Schemes for Content-Based Image Retrieval. In D. P. Huijsmans and A. W. M. Smeulders, editors, *Visual Information and Information Systems: Third International Conference (VISUAL'99)*, volume 1614 of *Lecture Notes in Computer Science (LNCS)*, pages 549–556, Amsterdam, The Netherlands, June 2–4 1999. Springer.
- [427] D. M. Squire, W. Müller, H. Müller, and T. Pun. Content-based Query of Image Databases: Inspirations From Text Retrieval. *Pattern Recognition Letters (Selected Papers from The 11th Scandinavian Conference on Image Analysis SCIA '99)*, 21(13-14):1193–1198, 2000.
- [428] D. M. Squire, W. Müller, H. Müller, and J. Raki. Content-Based Query of Image Databases, Inspirations From Text Retrieval: Inverted Files, Frequency-Based Weights and Relevance Feedback. In *The 11th Scandinavian Conference on Image Analysis (SCIA '99)*, pages 143–149, Kangerlussuaq, Greenland, June 7–11 1999.

- [429] R. Sriram, J. M. Francos, and W. A. Pearlman. Texture Coding Using a Wold Decomposition Model. *IEEE Transactions on Image Processing*, 5(9):1382–1386, September 1996.
- [430] E. J. Stollnitz, T. D. DeRose, and D. H. Salesin. Wavelets for Computer Graphics: A Primer, Part 1. *IEEE Computer Graphics and Applications*, 15(3):76–84, May 1995.
- [431] E. J. Stollnitz, T. D. DeRose, and D. H. Salesin. Wavelets for Computer Graphics: A Primer, Part 2. *IEEE Computer Graphics and Applications*, 15(4):75–85, July 1995.
- [432] M. A. Stricker and A. Dimai. Color Indexing with Weak Spatial Constraints. In I. K. Sethi and R. C. Jain, editors, *Storage and Retrieval for Image and Video Databases IV*, volume 2670 of *SPIE Proceedings*, pages 29–40, San Jose, CA, USA, March 1996.
- [433] M. A. Stricker and M. Orengo. Similarity of Color Images. In W. Niblack and R. C. Jain, editors, *Storage and Retrieval for Image and Video Databases III*, volume 2420 of *SPIE Proceedings*, pages 381–392, San Jose, CA, USA, March 1995.
- [434] D. Sutanto and C. H. C. Leung. Automatic Index Expansion For Concept-Based Image Query. In D. P. Huijsmans and A. W. M. Smeulders, editors, *Visual Information and Information Systems: Third International Conference (VISUAL'99)*, volume 1614 of *Lecture Notes in Computer Science (LNCS)*, pages 399–408, Amsterdam, The Netherlands, June 2–4 1999. Springer.
- [435] I. Sutherland. *Sketchpad: a man-machine graphical communication system*. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, USA, January 1963.
- [436] M. J. Swain and D. H. Ballard. Color Indexing. *International Journal of Computer Vision*, 7(1):11–32, November 1991.

- [437] A. Tam and C. H. C. Leung. Structured Natural-Language Descriptions for Semantic Content Retrieval of Visual Materials. *Journal of the American Society for Information Science and Technology*, 52(11):930–937, 2001.
- [438] H. Tamura, S. Mori, and T. Yamawaki. Textural Features Corresponding to Visual Perception. *IEEE Transactions on Systems, Man and Cybernetics*, SMC-8:460–473, June 1978.
- [439] H. Tamura and N. Yokoya. Image Database Systems: A Survey. *Pattern Recognition*, 17(1):29–43, 1984.
- [440] P. M. Tardif and A. Zaccarin. Multiscale Autoregressive Image Representation for Texture Segmentation. In E. R. Dougherty and J. T. Astola, editors, *Nonlinear Image Processing VIII*, volume 3026, pages 327–337, San Jose, CA, USA, 1997. SPIE.
- [441] L. Taycher, M. La Cascia, and S. Sclaroff. Image Digestion and Relevance Feedback in the ImageRover WWW Search Engine. In *Proceedings of the Second International Conference On Visual Information Systems (VISUAL'97)*, pages 85–94, San Diego, CA, USA, December 15–17 1997. Knowledge Systems Institute.
- [442] C.-H. Teh and R. T. Chin. On Image Analysis by the Methods of Moments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10(4):496–513, July 1988.
- [443] Q. Tian, N. Sebe, E. Louprias, M. S. Lew, and T. S. Huang. Image Retrieval Using Wavelet-based Salient Points. *Journal of Electronic Imaging*, 10(4):1132–1141, October 2001.
- [444] K. Tieu and P. Viola. Boosting Image Retrieval. In *Proceedings of the 2000 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2000)*, pages 228–235, Hilton Head Island, SC, USA, June 13–15 2000. IEEE Computer Society.

- [445] S. Tong and E. Chang. Support vector machine active learning for image retrieval. In *Proceedings of the 9th ACM International Conference on Multimedia (ACM Multimedia 2001)*, pages 107–118, Ottawa, Canada, September 30 - October 5 2001. ACM Press.
- [446] R. S. Torres, C. G. Silva, C. B. Medeiros, and H. V. Rocha. Visual structures for image browsing. In *Proceedings of the Twelfth International Conference on Information and Knowledge Management (CIKM '03)*, pages 49–55, New Orleans, LA, USA, November 3–8 2003. ACM Press.
- [447] C. Tsinaraki, P. Polydoros, N. Moutoutzis, and S. Christodoulakis. Coupling OWL with MPEG-7 and TV-Anytime for Domain-specific Multimedia Information Integration and Retrieval. In *Recherche d'Informations Assistée par Ordinateur (RIAO'2004) Computer-Assisted Information Retrieval*, Avignon, France, April 26–28 2004.
- [448] Z. Tu and S.-C. Zhu. Image Segmentation by Data-Driven Markov Chain Monte Carlo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):657–673, May 2002.
- [449] M. Tănase-Avătavului. *Shape Decomposition and Retrieval: Vorm Decompositie en het Opzoeken van Figuren*. PhD thesis, Institute of Information and Computing Sciences, Utrecht University, Utrecht, The Netherlands, February 2005.
- [450] M. Tuceryan and A. K. Jain. Texture Analysis. In C. H. Chen and L. F. Pau, editors, *The Handbook of Pattern Recognition and Computer Vision (2nd Edition)*, pages 207–248, River Edge, NJ, USA, 1998. World Scientific Publishing Co., Inc.
- [451] G. Tzagkarakis, B. Beferull-Lozano, and P. Tsakalides. Rotation-Invariant Texture Retrieval With Gaussianized Steerable Pyramids. *IEEE Transactions on Image Processing*, 15(9):2702–2718, September 2006.

- [452] A. Utenpattanant, O. Chitsobhuk, and A. Khawne. Color Descriptor for Image Retrieval in Wavelet Domain. In *The 8th International Conference on Advanced Communication Technology (ICACT 2006)*, pages 818–821, Phoenix Park, Korea, February 20–22 2006.
- [453] I. Valova, B. Rachev, and M. Vassilakopoulos. Optimization of the Algorithm for Image Retrieval by Color Features. In *International Conference on Computer Systems and Technologies (CompSysTech'2006)*, pages II.17.1 – II.17.5, University of Veliko Tarnovo, Bulgaria, June 15–16 2006.
- [454] C. J. van Rijsbergen. *Information Retrieval*. Prentice Hall, Englewood Cliffs, New Jersey, USA, 1979.
- [455] N. Vasconcelos and A. Lippman. A probabilistic architecture for content-based image retrieval. In *Proceedings of the 2000 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2000)*, pages 216–221, Hilton Head Island, SC, USA, June 13–15 2000. IEEE Computer Society.
- [456] N. Vasconcelos and A. Lippman. Learning over multiple temporal scales in image databases. In D. Vernon, editor, *6th European Conference on Computer Vision (ECCV 2000)*, volume 1842 of *Lecture Notes in Computer Science (LNCS)*, pages 33–47, Dublin, Ireland, June 26–30 2000. Springer.
- [457] N. Vasconcelos and A. Lippmann. A Unifying View of Image Similarity. In A. Sanfeliu, J. J. Villanueva, M. Vanrell, R. Alc azar, J.-O. Eklundh, and Y. Aloimonos, editors, *Proceedings of the 15th International Conference on Pattern Recognition (ICPR'00)*, pages 1–4, Barcelona, Spain, September 3–8 2000. IEEE Computer Society.
- [458] A. Vellaikal and C.-C. J. Kuo. Hierarchical Clustering Techniques for Image Database Organization and Summarization. In C.-C. J. Kuo, S.-F. Chang, and S. Panchanathan, editors, *Multimedia Storage and Archiving Systems III*,

volume 3527 of *SPIE Proceedings*, pages 68–79, San Jose, CA, USA, October 1998.

- [459] R. C. Veltkamp. Shape Matching: Similarity Measures and Algorithms. In *Proceedings of the SMI 2001 International Conference on Shape Modelling and Applications*, pages 188–197, Genova, Italy, May 07–11 2001.
- [460] R. C. Veltkamp and L. J. Latecki. Properties and Performances of Shape Similarity Measures. In V. Batagelj, A. Ferligoj, and A. Žiberna, editors, *Proceedings of the IFCS06 Conference on Data Science and Classification*, pages 47–56, Ljubljana, Slovenia, July 25–29 2006. Springer.
- [461] R. C. Veltkamp and M. Tanase. Content-Based Image Retrieval Systems: A Survey (revised and extended version of Technical Report UU-CS-2000-34). Technical report, Department of Computer Science, Utrecht University, Utrecht, The Netherlands, 2002.
- [462] J. Vendrig, M. Worring, and A. W. M. Smeulders. Filter Image Browsing - Exploiting Interaction in Image Retrieval. In D. P. Huijsmans and A. W. M. Smeulders, editors, *Visual Information and Information Systems: Third International Conference (VISUAL'99)*, volume 1614 of *Lecture Notes in Computer Science (LNCS)*, pages 147–154, Amsterdam, The Netherlands, June 2–4 1999. Springer.
- [463] C. C. Venters and M. Cooper. Content-Based Image Retrieval. Technical Report JTAP-054, JISC Technology Application Program, University of Manchester, Manchester, UK, 2000.
- [464] L. von Ahn and L. Dabbish. Labeling Images With a Computer Game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '04)*, pages 319–326, Vienna, Austria, April 24–29 2004. ACM Press.

- [465] E. M. Voorhees. Overview of the Seventh Text REtrieval Conference (TREC-7). In E. M. Voorhees and D. K. Harman, editors, *Proceedings of the Seventh Text REtrieval Conference (TREC-7)*, pages 1–24, Gaithersburg, MD, USA, November 9–11 1998. Department of Commerce, National Institute of Standards and Technology.
- [466] E. M. Voorhees. Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '98)*, pages 315–323, Melbourne, Australia, August 24–28 1998. ACM Press.
- [467] E. M. Voorhees. Evaluation by Highly Relevant Documents. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2001)*, pages 74–82, New Orleans, LA, USA, September 9–13 2001. ACM Press.
- [468] E. M. Voorhees, editor. *Proceedings of the Tenth Text REtrieval Conference (TREC 2001)*, Gaithersburg, MD, USA, November 13–16 2001. Department of Commerce, National Institute of Standards and Technology.
- [469] E. M. Voorhees. The philosophy of information retrieval information. In C. Peters, M. Braschler, J. Gonzalo, and M. Kluck, editors, *Evaluation of Cross-Language Information Retrieval Systems: Second Workshop of the Cross-Language Evaluation Forum (CLEF 2001)*, volume 2406 of *Lecture Notes in Computer Science (LNCS)*, pages 355–370, Darmstadt, Germany, September 3–4 2002. Springer.
- [470] E. M. Voorhees. Overview of the TREC 2004 Robust Retrieval Track. In E. M. Voorhees and L. P. Buckland, editors, *Proceedings of the Thirteenth Text REtrieval Conference (TREC 2004)*, pages 45–49, Gaithersburg, MD, USA, November 16–19 2004.

- [471] E. M. Voorhees. The TREC Robust Retrieval Track. *SIGIR Forum*, 39(1):11–20, 2005.
- [472] E. M. Voorhees. The TREC 2005 robust track. *SIGIR Forum*, 40(1):41–48, 2006.
- [473] E. M. Voorhees and C. Buckley. The Effect of Topic Set Size on Retrieval Experiment Error. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2002)*, pages 316–323, Tampere, Finland, August 11–15 2001. ACM Press.
- [474] E. M. Voorhees and D. K. Harman, editors. *Proceedings of the Seventh Text REtrieval Conference (TREC-7)*, Gaithersburg, MD, USA, November 9–11 1998. Department of Commerce, National Institute of Standards and Technology.
- [475] E. M. Voorhees and D. K. Harman. Overview of the Sixth Text REtrieval Conference (TREC-6). *Information Processing and Management*, 36(1):3–35, 2000.
- [476] H. Voorhees and T. Poggio. Computing Texture Boundaries From Images. *Nature*, 333:364–367, May 1988.
- [477] J. W. Wallis, M. M. Miller, T. R. Miller, and T. H. Vreeland. An Internet-based nuclear medicine teaching file. *The Journal of Nuclear Medicine*, 36(8):1520–1527, 1995.
- [478] H. Wang, F. Guo, D. D. Feng, and J. S. Jin. A Signature for Content-Based Image Retrieval Using a Geometrical Transform. In *Proceedings of the Sixth ACM International Conference on Multimedia (MULTIMEDIA '98)*, pages 229–234, Bristol, UK, September 12–16 1998. ACM Press.

- [479] J. Z. Wang, J. Li, R. M. Gray, and G. Wiederhold. Unsupervised Multiresolution Segmentation for Images With Low Depth of Field. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(1):85–90, January 2001.
- [480] J. Z. Wang, J. Li, and G. Wiederhold. SIMPLIcity: Semantics–Sensitive Integrated Matching for Picture LIbraries. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(9):1–17, September 2001.
- [481] J. Z. Wang, G. Wiederhold, O. Firschein, and S. Xin Wei. Wavelet–Based Image Indexing Techniques with Partial Sketch Retrieval Capability. In *Proceedings of the Fourth Forum on Research and Technology Advances in Digital Libraries (ADL '97)*, pages 13–24, Washington, DC, USA, May 7–9 1997. IEEE Computer Society.
- [482] T. P. Weldon and W. E. Higgins. Integrated Approach to Texture Segmentation Using Multiple Gabor Filters. In P. Delogne, editor, *Proceedings of the 1996 International Conference on Image Processing (ICIP'96)*, pages 955–958, Lausanne, Switzerland, September 16–19 1996.
- [483] T. Westerveld. Image Retrieval: Content versus Context. In *Recherche d'Informations Assistée par Ordinateur (RIAO'2000) Computer–Assisted Information Retrieval*, volume 1, pages 276–284, Paris, France, April 12–14 2000.
- [484] T. Westerveld and R. van Zwol. Benchmarking Multimedia Search in Structured Collections. In N. Fuhr, M. Lalmas, and A. Trotman, editors, *INEX 2006 Workshop Pre-Proceedings*, pages 313–320, Schloss Dagstuhl, Germany, December 18–20 2006. DELOS: Network of Excellence on Digital Libraries.
- [485] T. Westerveld and R. van Zwol. INEX REPORT: Multimedia Retrieval at INEX 2006. *SIGIR Forum*, 41(1), to appear 2007.
- [486] T. Westerveld and R. van Zwol. The INEX 2006 Multimedia Track. In N. Fuhr, M. Lalmas, and A. Trotman, editors, *Advances in XML Information*

Retrieval: Fifth International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2006, Lecture Notes in Computer Science/Lecture Notes in Artificial Intelligence (LNCS/LNAI), Schloss Dagstuhl, Germany, to appear 2007. Springer.

- [487] D. A. White and R. C. Jain. Algorithms and Strategies for Similarity Retrieval. Technical Report VCL-96-101, Visual Computing Laboratory, University of California, San Diego, La Jolla, CA, USA, July 1996.
- [488] D. A. White and R. C. Jain. Similarity Indexing: Algorithms and Performance. In I. K. Sethi and R. C. Jain, editors, *Storage and Retrieval for Image and Video Databases IV*, volume 2670 of *SPIE Proceedings*, pages 62–73, San Jose, CA, USA, March 1996.
- [489] T. Wilhelm and M. Eibl. ImageCLEF 2006 Experiments at the Chemnitz Technical University. In *CLEF working notes*, Alicante, Spain, September 2006.
- [490] I. H. Witten, A. Moffat, and T. C. Bell. *Managing Gigabytes: Compressing and Indexing Documents and Images*. Van Nostrand Reinhold, 115 Fifth Avenue, New York, NY 10003, USA, 1994.
- [491] C. Wolf and J.-M. Jolion. Object count/area graphs for the evaluation of object detection and segmentation algorithms. *International Journal on Document Analysis and Recognition*, 8(4):280–296, August 2006.
- [492] C. Womser-Hacker. Multilingual Topic Generation within the CLEF 2001 Experiments. In C. Peters, M. Braschler, J. Gonzalo, and M. Kluck, editors, *Evaluation of Cross-Language Information Retrieval Systems: Second Workshop of the Cross-Language Evaluation Forum (CLEF 2001)*, volume 2406 of *Lecture Notes in Computer Science (LNCS)*, pages 389–393, Darmstadt, Germany, September 3–4 2002. Springer.

- [493] M. Worring, A. W. M. Smeulders, and S. Santini. Interaction in Content-Based Image Retrieval: An Evaluation of the State of the Art. In R. Laurini, editor, *Fourth International Conference On Visual Information Systems (VISUAL'2000)*, volume 1929 of *Lecture Notes in Computer Science (LNCS)*, pages 26–36, Lyon, France, November 2–4 2000. Springer.
- [494] P. Wu, B. S. Manjunath, S. D. Newsam, and H. D. Shin. A Novel Texture Descriptor for Image Retrieval and Browsing. In *IEEE Workshop on Content-based Access of Image and Video Libraries (CBAIVL'99)*, pages 3–7, Fort Collins, CO, USA, June 22 1999. IEEE Computer Society.
- [495] X. Xie, H. Liu, S. Goumaz, and W.-Y. Ma. Learning User Interest For Image Browsing on Small-Form-Factor Devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '05)*, pages 671–680, Portland, OR, USA, April 2–7 2005. ACM Press.
- [496] K. Yanai, N. V. Shirahatti, P. Gabbur, and K. Barnard. Evaluation Strategies for Image Understanding and Retrieval. In *Proceedings of the Seventh ACM SIGMM International Workshop on Multimedia Information Retrieval (MIR'05)*, pages 217–226, Singapore, November 10–11 2005. ACM Press.
- [497] L. Yang and F. Albrechtsen. Fast Computation of Invariant Geometric Moments: A New Method Giving Correct Results. In *Proceedings of the 12th IAPR International Conference on Pattern Recognition (Volume 1)*, pages 201–204, Jerusalem, Israel, October 9–13 1994.
- [498] A. Yavlinsky, E. Schofield, and S. M. Rüger. Automated Image Annotation Using Global Features and Robust Nonparametric Density Estimation. In W. K. Leow, M. S. Lew, T.-S. Chua, W.-Y. Ma, L. Chaisorn, and E. M. Bakker, editors, *Proceedings of the 4th International Conference on Image and Video Retrieval (CIVR 2005)*, volume 3568 of *Lecture Notes in Computer Science (LNCS)*, pages 507–517, Singapore, July 20–22 2005. Springer.

- [499] H. Ye and G. Xu. Fast Search in Large-Scale Image Database Using Vector Quantization. In *Proceedings of the Second International Conference on Image and Video Retrieval (CIVR 2003)*, volume 2728 of *Lecture Notes in Computer Science (LNCS)*, pages 477–487, Urbana, IL, USA, July 24–25 2003. Springer.
- [500] E. Yom-Tov, S. Fine, D. Carmel, and A. Darlow. Learning to Estimate Query Difficulty: Including Applications to Missing Content Detection and Distributed Information Retrieval. In G. Marchionini, A. Moffat, J. Tait, R. Baeza-Yates, and N. Ziviani, editors, *Proceedings of the Twenty-Eighth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2005)*, pages 512–519, Salvador, Brazil, August 15–19 2005. ACM Press.
- [501] C. Zhai and J. Lafferty. A Study of Smoothing Methods for Language Models Applied to Information Retrieval. *ACM Transactions on Information Systems*, 22(2):179–214, April 2004.
- [502] H. J. Zhang and D. Zhong. A Scheme for Visual Feature-Based Image Indexing. In W. Niblack and R. C. Jain, editors, *Storage and Retrieval for Image and Video Databases III*, volume 2420 of *SPIE Proceedings*, pages 36–46, San Jose, CA, USA, March 1995.
- [503] V. Zhang, B. Rey, E. Stipp, and R. Jones. Geomodification in Query Rewriting. In *Online Proceedings of the Third Workshop on Geographic Information Retrieval at SIGIR 2006 (GIR'06)*, Seattle, WA, USA, August 10 2006.
- [504] D. Zhou. *Texture Analysis and Synthesis Using a Generic Markov-Gibbs Image Model*. PhD thesis, The University of Auckland, Auckland, New Zealand, February 2006.
- [505] X. S. Zhou and T. S. Huang. Relevance Feedback in Image Retrieval: A Comprehensive Review. *Multimedia Systems*, 8(6):536–544, 2003.

- [506] Z.-H. Zhou, K.-J. Chen, and H.-B. Dai. Enhancing relevance feedback in image retrieval using unlabeled data. *ACM Transactions on Information Systems (TOIS)*, 24(2):219–244, April 2006.
- [507] L. Zhu, A. Rao, and A. Zhang. Theory of keyblock-based image retrieval. *ACM Transactions on Information Systems (TOIS)*, 20(2):224–257, 2002.
- [508] L. Zhu, C. Tang, A. Rao, and A. Zhang. Using Thesaurus to model keyblock-based image retrieval. In *Proceedings of the Second International Conference on Multimedia and Exposition (ICME'2001)*, pages 237–240, Tokyo, Japan, August 22–25 2001. IEEE Computer Society.
- [509] S. Zinger, C. Millet, B. Mathieu, G. Grefenstette, P. Hède, and P.-A. Moëllic. Extracting an Ontology of Portrayable Objects from WordNet. In A. Hanbury and H. Müller, editors, *MUSCLE/ImageCLEF Workshop on Image and Video Retrieval Evaluation*, pages 17–24, Vienna, Austria, September 20 2005. MUSCLE Network of Excellence.
- [510] J. Zobel. How reliable are the results of large-scale information retrieval experiments? In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '98)*, pages 307–314, Melbourne, Australia, August 24–28 1998. ACM Press.
- [511] R. v. Zwol, G. Kazai, and M. Lalmas. INEX 2005 Multimedia Track. In N. Fuhr, M. Lalmas, S. Malik, and G. Kazai, editors, *Advances in XML Information Retrieval and Evaluation*, volume 3977 of *Lecture Notes in Computer Science (LNCS)*, pages 497–510, Dagstuhl, Germany, November 28–30 2006. Springer.