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

\[
t \in \{ \text{numerals}, \text{nouns}, \text{verbs}, \text{adjectives}, \text{adverbs}, \text{adjuncts} \} \quad (5.1)
\]

where adjuncts are either locative, temporal, causal or modificative adjuncts.
Definition 5.2.2 Given a sentence $T$, we define the topic sentence $\mathcal{T}$ as

$$\mathcal{T} := \{t_1, t_2, \ldots, t_K\}$$  \hfill (5.2)

where $t_k$ denotes the $k^{th}$ topic sentence element$^1$ 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

$$\mathcal{R} := \{\mathcal{R}_1, \mathcal{R}_2, \ldots, \mathcal{R}_K\},$$  \hfill (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.

$^1$We 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 $R$ denote the set of relevant images for a query term, and $N_D$ 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 $\alpha$ as

$$\alpha := 1 - \frac{|N_D \cap R|}{|R|}$$

(5.4)

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

The definition of $\alpha$ 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 ($|R| = 5000$). We further assume that 3000 of these images are directly retrieved because they are annotated with the term “people” ($|N_D| = 3000$), 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

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

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 $R$ the set of relevant images for a topic term, and $N$ the set of images retrieved, we define the factor for element ambiguity $\beta$ as

$$\beta := 1 - \frac{|R|}{|N \cup R|}$$

(5.5)
The definition of $\beta$ 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 ($|N| = 1000$), although only 400 are, in fact, relevant ($|R| = 400$). Thus,

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

**Definition 5.2.6** Let $\alpha$ and $\beta$ be as defined in Definition 5.2.4 and Definition 5.2.5 respectively. We define the annotation gap factor $\gamma$ as

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

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 $\eta$ and $\theta$.](image)
5.2.3 Topic Difficulty

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

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

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

We also denote $\mathcal{R}_0^* = I$ 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,

$$\mathcal{R}_1 := \{\mathcal{R}_{1,1}, \mathcal{R}_{1,2}, \ldots, \mathcal{R}_{1,K}\},$$  \hfill (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 $\mathcal{R}_{1,k} \ (k \in \{1, \ldots, K\})$ 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,

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

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

where $\gamma_1^*$ 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 $k - 1$):

$$\mathcal{R}_j := \{\mathcal{R}_{j,1}, \mathcal{R}_{j,2}, \ldots, \mathcal{R}_{j,K-j+1}\},$$

(5.11)

where

$$\mathcal{R}_{j,k} := \mathcal{R}_{j-1,k} \cap \mathcal{R}_{j-1,k}.$$  

(5.12)

Now, we define

$$\mathcal{R}_j^* := \mathcal{R}_{j,l},$$

(5.13)

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

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

(5.14)

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

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

(5.15)

where $\gamma_j^*$ 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

$$d := \sum_{j=1}^{K} d_j.$$  

(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 $T$ which contains two topic sentence elements: the noun $\text{people} (t_1)$ and the locative adjunct $\text{San Francisco} (t_2)$, whereby we remove the preposition ($\text{in}$) because it is found in most stop-word lists:

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

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 $\text{people} (t_1)$ and $\mathcal{R}_2$ the set of all the images that were taken in $\text{San Francisco} (t_2)$. 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 $\mathcal{T}_1 = T$ and $\mathcal{R}_{1,k} = \mathcal{R}_{0,k}$ by definition, and therefore

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

$$\mathcal{R}_1 = \{\mathcal{R}_{1,1}, \mathcal{R}_{1,2}\}.$$
We first compare the cardinalities of both sets $R_{1,1}$ and $R_{1,2}$: there are 5000 images in the database that contain people ($|R_{1,1}| = 5000$) and 400 images that were taken in San Francisco ($|R_{1,2}| = 400$). The most significant topic element is therefore San Francisco, because it contains the minimum number of relevant images in the first iteration:

$$R^*_1 = R_{1,2}.$$ 

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, $R^*_0 = I$, and therefore

$$df(R^*_1) = \frac{|R_{1,2}|}{|R^*_0|} = \frac{|R_{1,2}|}{|I|} = \frac{400}{20000} = 0.02.$$ 

Then, we compute the annotation gap factor for the most significant element of the first iteration ($\gamma^*_1$) using the values that we have already determined in Section 5.2.2 ($\alpha = 0$, $\beta = 0.6$, $\eta = 1.2$ and $\theta = 0.6$):

$$\gamma^*_1 = \eta + [\theta \alpha^*_1 + (1 - \theta)\beta^*_1] = 1.2 + [0.6 \times 0.4 + 0.4 \times 0.6] = 1.68.$$ 

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

$$d_1 = 1 - df(R^*_1) \gamma^*_1 = 1 - 0.02 \times 1.68 = 1.65.$$ 

**Iteration 2.** In the second iteration ($j = 2$), we first have to build $T_2$ and $R_2$ by intersecting the set of all relevant images of the most significant topic element of the previous iteration $R^*_1$ with all the sets of relevant images of the previous iteration $R_{1,k}$, except $R^*_1 = R_{1,2}$ itself:

$$T_2 = \{(people, SanFrancisco)\},$$

$$R_2 = R_{2,1} = \{R^*_1 \cap R_{1,1}\}.$$ 

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

$$R^*_2 = R_{2,1}.$$
We now assume that 200 images show people in the Californian city of San Francisco (|R_2^*| = 200), that 150 of them are found by direct hits (|N_{D_2}^* \cap 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 (|N_2^* \cup R_2^*| = 300). The linear document frequency df therefore is
\[ df(R_2^*) = \frac{|R_{2,1}|}{|R_2^*|} = \frac{|\{\text{images of people in San Francisco}\}|}{|\{\text{images of San Francisco}\}|} = \frac{200}{400} = 0.5, \]
the factor for vocabulary mismatch and incomplete and incorrect annotation
\[ \alpha_2^* = 1 - \frac{|N_{D_2}^* \cap R_2^*|}{|R_2^*|} = 1 - \frac{150}{200} = 1 - 0.75 = 0.25, \]
the factor for word ambiguity
\[ \beta_2^* = 1 - \frac{|R_2^*|}{|N_2^* \cup R_2^*|} = 1 - \frac{200}{300} = 1 - 0.66 = 0.33, \]
the factor for the annotation gap
\[ \gamma_2^* = \eta + [\theta \alpha_2^* + (1 - \theta) \beta_2^*] = 1.2 + [0.6 \times 0.25 + 0.4 \times 0.33] = 1.38, \]
and the difficulty for the second iteration
\[ d_2 = [1 - df(R_2^*)] \times \gamma_2^* = [1 - 0.5] \times 1.38 = 0.69. \]

**Final result.** The total topic difficulty of topic T (“people in San Francisco”) for an image collection I after two iterations amounts to
\[ d(T, I) = \sum_{j=1}^{2} d_j = 1.65 + 0.69 = 2.34. \]

**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 I that contains a total of 20,000 images (N = |I| = 20,000).
Initial condition. We first define the topic sentence $T$ 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:

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

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 $T_1 = T$ and $\mathcal{R}_{1,k} = \mathcal{R}_{0,k}$ by definition, and therefore

$$T_1 = \{\text{photo, female, guide}\},$$

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

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, $R^*_0 = T$, and therefore

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

Then, we compute the annotation gap factor for the most significant element of the first iteration ($\gamma^*_1$). In the IAPR TC-12 image collection, all photos of guides are also annotated as such ($|N^*_0 \cap R^*_1| = 90$) and there are no other photos of guides in the database that are not relevant ($|N^*_1 \cup R^*_1| = 90$), therefore $\alpha^*_1 = 0$ and $\beta^*_1 = 0$, and subsequently

$$\gamma^*_1 = \eta + \left[\theta \alpha^*_1 + (1 - \theta) \beta^*_1\right] = 1.2 + [0.6 \times 0 + 0.4 \times 0] = 1.2.$$

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

$$d_1 = \left[1 - df(R^*_1)\right] \gamma^*_1 = \left[1 - 0.0045\right] \times 1.2 = 1.194.$$

**Iteration 2.** In the second iteration ($j = 2$), we again first have to build $T_2$ and $R_2$ by intersecting the set of all relevant images of the most significant topic element of the previous iteration $R^*_1$ with all the sets of relevant images of the previous iteration $R^*_1$, except $R^*_1 = R^*_{1,3}$ itself:

$$T_2 = \{(\text{guide, photo}), (\text{guide, female})\},$$

$$R_2 = \{R_{2,1}, R_{2,2}\} = \{R^*_1 \cap R_{1,1}, R^*_1 \cap R_{1,2}\}.$$

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

$$R^*_2 = R_{2,2},$$

with a conditional linear document frequency $df$ of:

$$df(R^*_2) = \frac{|R^*_{2,2}|}{|R^*_1|} = \frac{|\left\{\text{images of female guides}\right\}|}{|\left\{\text{images of guides}\right\}|} = \frac{26}{90} = 0.289.$$
A search for female guides would return 5 direct and relevant hits from the collection \(|\mathcal{N}_D^* \cap \mathcal{R}_1^*| = 5\), and there are 26 images that are directly or indirectly annotated with female guides \(|\mathcal{N}_I^* \cup \mathcal{R}_1^*| = 26\). Thus, the factor for vocabulary mismatch and incomplete and incorrect annotation is

\[
\alpha_2^* = 1 - \frac{|\mathcal{N}_D^* \cap \mathcal{R}_2^*|}{|\mathcal{R}_2^*|} = 1 - \frac{5}{26} = 1 - 0.19 = 0.81,
\]

the factor for word ambiguity

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

the factor for the annotation gap

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

and the difficulty for the second iteration

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

**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 ($|N_{D1}^* \cap R_1^*| = 0$), while it contains 26 images that are directly or indirectly annotated with photos of female guides ($|N_1^* \cup R_1^*| = 26$). Thus, the factor for vocabulary mismatch and incomplete and incorrect annotation in the third iteration is
\[ \alpha_3^* = 1 - \frac{|N_{D3}^* \cap R_3^*|}{|R_3^*|} = 1 - \frac{0}{26} = 1 - 0 = 1, \]
the factor for word ambiguity
\[ \beta_3^* = 1 - \frac{|R_3^*|}{|N_3^* \cup R_3^*|} = 1 - \frac{26}{26} = 1 - 1 = 0, \]
the factor for the annotation gap
\[ \gamma_3^* = \eta + [\theta \alpha_3^* + (1 - \theta) \beta_3^*] = 1.2 + [0.6 \times 1 + 0.4 \times 0] = 1.8, \]
and the difficulty for the third iteration
\[ d_3 = \left[1 - df(R_3^*)\right] \times \gamma_3^* = \left[1 - 1\right] \times 1.686 = 0. \]

**Final result.** The total topic difficulty of topic $T$ (“photos of female guides”) for the IAPR TC-12 image collection $I$ after three iterations amounts to
\[ d(T, I) = 3 \sum_{j=1}^3 d_j = 1.194 + 1.199 + 0 = 2.393. \]

### 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).](image-url)
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 $\sigma_X$ and $\sigma_Y$

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

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 $\rho$; thus, we aim for a strong negative correlation: the higher the topic difficulty measure, the lower the precision values for the results.

<table>
<thead>
<tr>
<th>Year</th>
<th>#runs</th>
<th>#topics</th>
<th>$\rho(d, \text{MAP})$</th>
<th>$\rho(d, P_{10})$</th>
<th>$\rho(d, P_{20})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2004</td>
<td>30</td>
<td>25</td>
<td>-0.818</td>
<td>-0.692</td>
<td>-0.593</td>
</tr>
<tr>
<td>2005</td>
<td>62</td>
<td>28</td>
<td>-0.764</td>
<td>-0.671</td>
<td>-0.700</td>
</tr>
<tr>
<td>2006</td>
<td>40</td>
<td>60</td>
<td>-0.711</td>
<td>-0.647</td>
<td>-0.660</td>
</tr>
<tr>
<td>Total</td>
<td>132</td>
<td>113</td>
<td>-0.783</td>
<td>-0.686</td>
<td>-0.666</td>
</tr>
</tbody>
</table>

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 ($\gamma$). 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.

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

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 factor\textsuperscript{2}.

<table>
<thead>
<tr>
<th>Year</th>
<th>P1</th>
<th>IDF1</th>
<th>LSE</th>
<th>MSW</th>
<th>LSW</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>2006</td>
<td>-0.253</td>
<td>-0.346</td>
<td>-0.516</td>
<td>-0.490</td>
<td>-0.365</td>
<td>-0.647</td>
</tr>
<tr>
<td>2005</td>
<td>-0.371</td>
<td>-0.454</td>
<td>-0.504</td>
<td>-0.479</td>
<td>-0.380</td>
<td>-0.715</td>
</tr>
<tr>
<td>2004</td>
<td>0.017</td>
<td>-0.078</td>
<td>-0.492</td>
<td>-0.468</td>
<td>-0.395</td>
<td>-0.660</td>
</tr>
<tr>
<td>AVG</td>
<td>-0.202</td>
<td>-0.293</td>
<td>-0.504</td>
<td>-0.479</td>
<td>-0.380</td>
<td>-0.674</td>
</tr>
</tbody>
</table>

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 $\gamma$): using the algorithm as described in Section 5.2.3 with the probabilities (P2) or $tf-idf$ (IDF2) multiplied with average $\gamma$ for individual elements, the probabilities (P3) or $tf-idf$ (IDF3) multiplied with $\gamma$ at iterations, the sum of probabilities and $\gamma$ (P4), the sum of $td-idf$ and $\gamma$ (IDF4), and the probabilities (P5) or $tf-idf$ (IDF5) multiplied with $\gamma$ for individual elements.

\textsuperscript{2}We therefore only use IDF to denote the approaches using tf-idf weighting.
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 \textit{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.

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

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 \textit{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...
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, 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.

<table>
<thead>
<tr>
<th>Topic</th>
<th>d</th>
<th>MAP</th>
<th>P(10)</th>
<th>P(20)</th>
</tr>
</thead>
<tbody>
<tr>
<td>photos of radio telescopes</td>
<td>1.199</td>
<td>0.5914</td>
<td>0.5914</td>
<td>0.5350</td>
</tr>
<tr>
<td>animal statue</td>
<td>1.222</td>
<td>0.5974</td>
<td>0.7339</td>
<td>0.7113</td>
</tr>
<tr>
<td>Rome in April 1908</td>
<td>1.698</td>
<td>0.3840</td>
<td>0.4167</td>
<td>0.3583</td>
</tr>
</tbody>
</table>

Table 5.6: Examples of easy topics.

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

Table 5.7: Examples of difficult topics.
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 effec-
tiveness (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.

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

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, 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

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

Table 5.9: ImageCLEFphoto 2006 topics.
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 afore-mentioned 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 database\textsuperscript{1} 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 module\textsuperscript{2} 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

\textsuperscript{1}Version 4.1.20.

\textsuperscript{2}The primary keys of the tables are in bold letters, with the lines indicating the relationships between them.
also considers multilingual aspects, provides the possibility to include additional languages as well as alternative image representation types and formats.

The core of this design is made up of three tables: images, annotations and languages. Most of the other entities are the result of the normalisation process to avoid redundancies in the database and to guarantee the high flexibility and extensibility of the architecture. Moreover, three of these outsourced entities (objects, locations, countries) are linked with the table wn_synset, which itself provides the interface between the database of our benchmark administration system and that of WordNet – an ontology which hierarchically organises nouns, verbs and adjectives into synonym sets (synsets), each of them representing one underlying lexical concept (see Section 2.4.2).
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 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.

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 database\(^3\) to store the un-
derlying information, PHP\(^4\) for the implementation of the web-based user interface
and a Linux server\(^5\) 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.

---

\(^3\) Version 4.1.20.

\(^4\) PHP: Hypertext Preprocessor, Version 4.3.10-16.

\(^5\) Version 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 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 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.

![Figure 6.5: Edit authors.](image)

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

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.](image)

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.

![Topic overview](image)

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).
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
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.
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.
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.](image)

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).
6.3.4 Relevance Assessment

The blue link on the main page, *Relevance*, leads to the relevance assessment module of the benchmark administration system.\(^6\)

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.

\(^6\)This 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).

![Image of Relevance Assessment Overview](Image)

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”.

![Image of Relevance Assessment Page](Image)

Figure 6.21: Relevance assessment overview.
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 assessor (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 *FlickR*) 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.](image)

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.

```
<DOC>
<DOCNO>annotations/16/16019.xml</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>
```

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 setting 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
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/32/32647.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 ImageCLEF-photo 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.

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

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 < level ≤ 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 page\(^2\) 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 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 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.

<table>
<thead>
<tr>
<th>Relevance assessments</th>
<th>Average</th>
<th>Minimum</th>
<th>Median</th>
<th>Maximum</th>
<th>σ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pool size (images)</td>
<td>170.8</td>
<td>83</td>
<td>176</td>
<td>196</td>
<td>23.9</td>
</tr>
<tr>
<td>Total relevant (images)</td>
<td>100.3</td>
<td>22</td>
<td>69.5</td>
<td>419</td>
<td>100.0</td>
</tr>
<tr>
<td>Relevant through ISJ (in %)</td>
<td>77.5</td>
<td>31.6</td>
<td>83.6</td>
<td>98.0</td>
<td>18.2</td>
</tr>
</tbody>
</table>

Table 7.3: Relevance assessment statistics (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
(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

\(^3\)http://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 default\(^4\), 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].

\(^4\)Google, 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 results\(^5\). 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

\(^5\) These 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.
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 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, Miracle Daedalus University, Madrid, Spain)...

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

Table 7.5: Submission overview by dimensions.
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 monolingual English run, while four groups also submitted a total of eight monolingual German runs. The majority of runs was concerned with retrieval from English

<table>
<thead>
<tr>
<th>Query language</th>
<th>Caption language</th>
<th>Runs</th>
<th>Groups</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>English</td>
<td>49</td>
<td>11</td>
</tr>
<tr>
<td>Italian</td>
<td>English</td>
<td>15</td>
<td>4</td>
</tr>
<tr>
<td>Japanese</td>
<td>English</td>
<td>10</td>
<td>4</td>
</tr>
<tr>
<td>Simplified Chinese</td>
<td>English</td>
<td>10</td>
<td>3</td>
</tr>
<tr>
<td>French</td>
<td>English</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>Russian</td>
<td>English</td>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td>German</td>
<td>English</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>Spanish</td>
<td>English</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>Portuguese</td>
<td>English</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>Dutch</td>
<td>English</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>Traditional Chinese</td>
<td>English</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>Polish</td>
<td>English</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Visual</td>
<td>English</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>German</td>
<td>German</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>English</td>
<td>German</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>French</td>
<td>German</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Japanese</td>
<td>German</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Visual</td>
<td>(none)</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>Visual topics</td>
<td>(none)</td>
<td>6</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 7.6: Ad-hoc experiments listed by query and caption languages.
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 snapshot 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 \textit{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].

\section*{CELI}

The participants from \textit{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].

\section*{CINDI}

The CINDI group from \textit{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].

\section*{DCU}

\textit{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 \textit{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].
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 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

<table>
<thead>
<tr>
<th>Language (Captions)</th>
<th>Group</th>
<th>Run ID</th>
<th>MAP</th>
<th>P(20)</th>
<th>BPREF</th>
<th>GMAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>English (English)</td>
<td>CINDI</td>
<td>Exp</td>
<td>0.385</td>
<td>0.530</td>
<td>0.874</td>
<td>0.282</td>
</tr>
<tr>
<td>German (German)</td>
<td>NTU</td>
<td>DE-DE-AUTO-FB-TXTIMG</td>
<td>0.311</td>
<td>0.335</td>
<td>0.974</td>
<td>0.132</td>
</tr>
<tr>
<td>Portuguese (English)</td>
<td>NTU</td>
<td>PT-EN-AUTO-FB-TXTIMG</td>
<td>0.285</td>
<td>0.403</td>
<td>0.755</td>
<td>0.177</td>
</tr>
<tr>
<td>T. Chinese (English)</td>
<td>NTU</td>
<td>ZHS-EN-AUTO-FB-TXTIMG</td>
<td>0.279</td>
<td>0.464</td>
<td>0.669</td>
<td>0.154</td>
</tr>
<tr>
<td>Russian (English)</td>
<td>NTU</td>
<td>RU-EN-AUTO-FB-TXTIMG</td>
<td>0.279</td>
<td>0.408</td>
<td>0.755</td>
<td>0.153</td>
</tr>
<tr>
<td>Spanish (English)</td>
<td>NTU</td>
<td>SP-EN-AUTO-FB-TXTIMG</td>
<td>0.278</td>
<td>0.407</td>
<td>0.757</td>
<td>0.175</td>
</tr>
<tr>
<td>French (English)</td>
<td>NTU</td>
<td>FR-EN-AUTO-FB-TXTIMG</td>
<td>0.276</td>
<td>0.416</td>
<td>0.750</td>
<td>0.158</td>
</tr>
<tr>
<td>Visual (English)</td>
<td>NTU</td>
<td>AUTO-FB-TXTIMG</td>
<td>0.276</td>
<td>0.448</td>
<td>0.657</td>
<td>0.107</td>
</tr>
<tr>
<td>S. Chinese (English)</td>
<td>NTU</td>
<td>ZHS-EN-AUTO-FB-TXTIMG</td>
<td>0.272</td>
<td>0.392</td>
<td>0.750</td>
<td>0.168</td>
</tr>
<tr>
<td>Japanese (English)</td>
<td>NTU</td>
<td>JA-EN-AUTO-FB-TXTIMG</td>
<td>0.271</td>
<td>0.402</td>
<td>0.746</td>
<td>0.170</td>
</tr>
<tr>
<td>Italian (English)</td>
<td>NTU</td>
<td>IT-EN-AUTO-FB-TXTIMG</td>
<td>0.262</td>
<td>0.398</td>
<td>0.722</td>
<td>0.143</td>
</tr>
<tr>
<td>German (English)</td>
<td>DCU</td>
<td>combTextVisual_DEENEN</td>
<td>0.189</td>
<td>0.258</td>
<td>0.683</td>
<td>0.070</td>
</tr>
<tr>
<td>Dutch (English)</td>
<td>DCU</td>
<td>combTextVisual_NLENEN</td>
<td>0.184</td>
<td>0.234</td>
<td>0.640</td>
<td>0.063</td>
</tr>
<tr>
<td>English (German)</td>
<td>DCU</td>
<td>combTextVisual_ENDEEN</td>
<td>0.122</td>
<td>0.175</td>
<td>0.524</td>
<td>0.036</td>
</tr>
<tr>
<td>Polish (English)</td>
<td>Miracle</td>
<td>miratctdplena</td>
<td>0.108</td>
<td>0.139</td>
<td>0.428</td>
<td>0.005</td>
</tr>
<tr>
<td>French (German)</td>
<td>DCU</td>
<td>combTextVisual_FRDEEN</td>
<td>0.104</td>
<td>0.147</td>
<td>0.245</td>
<td>0.002</td>
</tr>
<tr>
<td>Visual (none)</td>
<td>RWTH</td>
<td>RWTH6-IFHTAM</td>
<td>0.063</td>
<td>0.182</td>
<td>0.366</td>
<td>0.022</td>
</tr>
<tr>
<td>Japanese (German)</td>
<td>NII</td>
<td>mcp.bl.jp_tger_2d_skl_dir</td>
<td>0.032</td>
<td>0.051</td>
<td>0.172</td>
<td>0.001</td>
</tr>
</tbody>
</table>

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

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

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

<table>
<thead>
<tr>
<th>Feedback</th>
<th>MAP</th>
<th>P(20)</th>
<th>BPREF</th>
<th>GMAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>With</td>
<td>0.1646 (0.0900)</td>
<td>0.2239 (0.1278)</td>
<td>0.5475 (0.2082)</td>
<td>0.0622 (0.0670)</td>
</tr>
<tr>
<td>Without</td>
<td>0.1277 (0.0548)</td>
<td>0.1816 (0.0693)</td>
<td>0.4761 (0.1525)</td>
<td>0.0309 (0.0363)</td>
</tr>
</tbody>
</table>

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 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 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

<table>
<thead>
<tr>
<th>Topics with...</th>
<th>avg MAP</th>
<th>avg ( P(20) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>directly taken from the log file</td>
<td>0.1296 (0.0928)</td>
<td>0.1987 (0.1335)</td>
</tr>
<tr>
<td>derived from the log file</td>
<td>0.1155 (0.0625)</td>
<td>0.1578 (0.0716)</td>
</tr>
<tr>
<td>not taken from the log file</td>
<td>0.2191 (0.1604)</td>
<td>0.2172 (0.1063)</td>
</tr>
</tbody>
</table>

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 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

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

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

<table>
<thead>
<tr>
<th>Topics rated as...</th>
<th>difficulty (d)</th>
<th>avg MAP</th>
<th>avg P(20)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(very) easy</td>
<td>0 ≤ d &lt; 2</td>
<td>0.4148 (0.1427)</td>
<td>0.4528 (0.1121)</td>
</tr>
<tr>
<td>medium</td>
<td>2 ≤ d &lt; 3</td>
<td>0.1742 (0.0875)</td>
<td>0.2454 (0.1215)</td>
</tr>
<tr>
<td>hard</td>
<td>3 ≤ d &lt; 4</td>
<td>0.1128 (0.0732)</td>
<td>0.1815 (0.1002)</td>
</tr>
<tr>
<td>very hard</td>
<td>d ≥ 4</td>
<td>0.0196 (0.0138)</td>
<td>0.0603 (0.0532)</td>
</tr>
</tbody>
</table>

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

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

Table 7.17: Topic overview by logical image representation completeness.

**7.4 Event Analysis and Evaluation**

_ImagelCLEFphoto 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 _ImagelCLEFphoto 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 _ImagelCLEFphoto 2007_.

309
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:

\[
ER = \frac{\sum \min(|A > B|, |B > A|)}{\sum (|A > B| + |A \equiv B| + |B > A|)}
\] (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:

\[
PT = \frac{\sum |A \equiv B|}{\sum (|A > B| + |A \equiv B| + |B > A|)}
\] (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%.
There is a consistent decrease in error rate as the fuzziness value increases, whereby \textit{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).

<table>
<thead>
<tr>
<th>ER for...</th>
<th>( f = 0.05 )</th>
<th>( f = 0.10 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAP</td>
<td>3.86 %</td>
<td>2.38 %</td>
</tr>
<tr>
<td>P(20)</td>
<td>5.92 %</td>
<td>3.97 %</td>
</tr>
<tr>
<td>GMAP</td>
<td>4.47 %</td>
<td>3.69 %</td>
</tr>
<tr>
<td>bpref</td>
<td>2.08 %</td>
<td>0.80 %</td>
</tr>
</tbody>
</table>

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.

Unsurprisingly, \textit{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).

\begin{table}[h]
\centering
\begin{tabular}{lcc}
\hline
 Difference & \textit{f = 0.05} & \textit{f = 0.10} \\
\hline
 MAP & 5.86 \% & 11.30 \% \\
 P(20) & 6.78 \% & 12.92 \% \\
 GMAP & 2.40 \% & 4.55 \% \\
 \textit{bpref} & 8.82 \% & 16.67 \% \\
\hline
\end{tabular}
\caption{Proportion of ties for fuzziness of 5\% and 10\%.}
\end{table}
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 $f = [0, 0.01, 0.02, ..., 0.20]$ and plot $PT$ and $ER$ in order to evaluate the stability of the collection (see Figure 7.10).

![Figure 7.10: ER-PT curves.](image)

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*.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Difference (f)</th>
<th>Ties</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAP</td>
<td>2.33 %</td>
<td>2.87 %</td>
</tr>
<tr>
<td>P(20)</td>
<td>7.14 %</td>
<td>9.42 %</td>
</tr>
<tr>
<td>GMAP</td>
<td>2.35 %</td>
<td>1.19 %</td>
</tr>
<tr>
<td>bpref</td>
<td>0.10 %</td>
<td>0.30 %</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Collection</th>
<th>Required fuzziness (f)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TREC-3</td>
<td>4.1%</td>
</tr>
<tr>
<td>TREC-4</td>
<td>5.5%</td>
</tr>
<tr>
<td>TREC-5</td>
<td>6.1%</td>
</tr>
<tr>
<td>TREC-6</td>
<td>7.2%</td>
</tr>
<tr>
<td>TREC-7</td>
<td>4.4%</td>
</tr>
<tr>
<td>TREC-8</td>
<td>4.3%</td>
</tr>
<tr>
<td>TREC-9</td>
<td>5.4%</td>
</tr>
<tr>
<td>NTCIR-3 (Chinese)</td>
<td>11.0%</td>
</tr>
<tr>
<td>NTCIR-3 (Japanese)</td>
<td>11.0%</td>
</tr>
<tr>
<td>NTCIR-3 (English)</td>
<td>14.0%</td>
</tr>
</tbody>
</table>

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
an 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.


<table>
<thead>
<tr>
<th>Year</th>
<th>avg d (std)</th>
<th>avg MAP (std)</th>
<th>avg P(20) (std)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2004</td>
<td>2.09 (0.37)</td>
<td>0.3720 (0.1701)</td>
<td>0.4001 (0.2382)</td>
</tr>
<tr>
<td>2005</td>
<td>2.36 (0.32)</td>
<td>0.3171 (0.1482)</td>
<td>0.3750 (0.1563)</td>
</tr>
<tr>
<td>2006</td>
<td>2.92 (0.69)</td>
<td>0.1584 (0.1157)</td>
<td>0.2280 (0.1392)</td>
</tr>
</tbody>
</table>

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 τ < 0.8 generally reflect noticeable changes in the rankings and suggest that the measures have a different evaluation emphasis, whereas correlations of τ > 0.9 can be considered as equivalent [467]. Hence, our set of performance measures would ideally exhibit correlations of τ < 0.9, because if any two measures had shown a correlation of τ > 0.9, one of them would have been redundant and could have been dropped as they both would have expressed the same evaluation emphasis.

<table>
<thead>
<tr>
<th>Kendall (τ)</th>
<th>MAP</th>
<th>P(20)</th>
<th>BPREF</th>
<th>GMAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAP</td>
<td>N/A</td>
<td>0.852</td>
<td>0.797</td>
<td>0.741</td>
</tr>
<tr>
<td>P(20)</td>
<td>0.852</td>
<td>N/A</td>
<td>0.735</td>
<td>0.742</td>
</tr>
<tr>
<td>BPREF</td>
<td>0.797</td>
<td>0.735</td>
<td>N/A</td>
<td>0.854</td>
</tr>
<tr>
<td>GMAP</td>
<td>0.741</td>
<td>0.742</td>
<td>0.854</td>
<td>N/A</td>
</tr>
</tbody>
</table>

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 τ < 0.9 (and most of them even τ < 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 $\text{MAP}$, $P(20)$, $\text{GMAP}$, and $\text{bpref}$ was also accepted by the majority of the groups, with all the participants agreeing on keeping $\text{MAP}$ as the leading measure to express retrieval performance. However, two participants expressed their concern about $P(20)$, and one participant questioned whether $\text{GMAP}$ and $\text{bpref}$ 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 alphanumerical 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 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 entries\(^1\), 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

\(^1\)As 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  Colombia 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)

\[ b_{pref} \] 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)

\[ D \] 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)

$\mathbf{d}_j$ Topic difficulty for iteration $j$ (defined in Equation 5.15)

$d(T, I)$ Topic difficulty of topic $T$ for image collection $I$ (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)

$H_{\text{max}}$ 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)

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

$I$ 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)

$N_O$ Number of objects in an image (described on Page 174)

$N_O(i)$ 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)

$N_{R_{\text{max}}}$ 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)
\( n_{rj} \) 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)

$\text{Rank}_1$ Rank of the first relevant image (introduced on Page 147)

$\text{Rank}_k$ Rank of the $k^{th}$ relevant image (introduced on Page 148)

$\text{Rank}_{rec}$ Reciprocal Rank of the first relevant image (defined in Equation 3.8)

$\overline{\text{Rank}}$ Average rank of relevant images (defined in Equation 3.9)

$S$ MUSCLE CIS Score (defined in Equation 3.14)

$T$ Topic sentence (defined in Equation 5.2)

$T_j$ 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)

$\alpha$ Factor for vocabulary mismatch and incomplete and incorrect annotation (defined in Equation 5.4)

$\beta$ Factor for element ambiguity (defined in Equation 5.5)

$\gamma$ Annotation gap factor (defined in Equation 5.6)

$\eta$ Tuning constant (introduced in Equation 5.6)
\( \theta \) Tuning constant (introduced in Equation 5.6)

\( \lambda \) Constant for calculation of \( GMAP \) (introduced in Equation 3.7)

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

\( \rho \) Object repetitiveness (defined in Equation 4.1)

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

\( \sigma_X \) Standard deviation of variable \( X \) (used in Equation 5.17)

\( \tau \) Kendall’s rank correlation coefficient (described on Page 317)
Bibliography


[137] J. Gonzalo, J. Karlgren, and P. D. Clough. iCLEF 2006 Overview: Searching
the FlickR 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–

of the American Society for Information Science (ASIS 1999), pages 665–674,

[139] A. A. Goodrum. Image Information Retrieval: An Overview of Current Re-
search. Informing Science, Special Issue on Information Science Research,

[140] C. C. Gotlieb and H. E. Kreyszig. Texture Descriptors Based on Co-
ocurrence Matrices. Computer Vision, Graphics, and Image Processing,

[141] V. Gouet and N. Boujemaa. On the Robustness of Color Points of Interest
for Image Retrieval. In Proceedings of the 2002 IEEE International Confer-
ence on Image Processing (ICIIP 2002), pages 377–380, Rochester, NY, USA,


[143] C. Grigorescu and N. Petkov. Distance Sets for Shape Filters and Shape
Recognition. IEEE Transactions on Image Processing, 12(10):1274–1286, Oc-
tober 2003.

368


369


[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-


[253] C. Luoni. Development of an Interface to a Database Storing the Features of a Multimedia Retrieval System. License thesis (BSc), Computer Vision and


Indexing and Retrieval in Image Databases. In L. W. MacDonald and M. R. 
Luo, editors, Color Imaging Science: Exploiting Digital Media. John Wiley, 

of the 1999 IEEE Conference on Computer Vision and Pattern Recognition 
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, edi-
tor, Proceedings of AH 2003: Workshop on Adaptive Hypermedia and Adap-
tive Web Based Systems, pages 217–235, Nottingham, UK, August 26–30 
2003.

Ontology-Based Photo Annotation. IEEE Intelligent Systems, 16(3):66–74, 


[390] S. Sclaroff, M. La Cascia, S. Sethi, and L. Taycher. Unifying Textual and Vi-


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, 


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


