# MINING PHOTOGRAPHIC COLLECTIONS TO ENHANCE THE PRECISION AND RECALL OF SEARCH RESULTS USING SEMANTICALLY CONTROLLED QUERY EXPANSION

Osama El Demerdash

A thesis

in

The Department

of

Computer Science and Software Engineering

Presented in Partial Fulfillment of the Requirements

For the Degree of Doctor of Philosophy

Concordia University

Montréal, Québec, Canada

April 2013

# CONCORDIA UNIVERSITY

## School of Graduate Studies

This is to certify that the thesis prepared

By: **Mr. Osama El Demerdash**

Entitled: **Mining Photographic Collections to Enhance the Precision and Recall of Search Results Using Semantically Controlled Query Expansion**

and submitted in partial fulfillment of the requirements for the degree of

**Doctor of Philosophy (Computer Science and Software Engineering)**

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the final examining commitee:

_____ Chair

_____ Dr. Diana Inkpen

_____ Dr. Peter Grogono

_____ Prof. PK Langshaw

_____ Dr. Sudhir Mudur

_____ Dr. Leila Kosseim

_____ Dr. Sabine Bergler

Approved _____

Chair of Department or Graduate Program Director

_____ 20 _____ _____

Dr. Robin Drew, Dean

Faculty of Engineering and Computer Science

# Abstract

Mining Photographic Collections to Enhance the Precision and Recall of Search Results Using Semantically Controlled Query Expansion

Osama El Demerdash, Ph.D.

Concordia University, 2013

Driven by a larger and more diverse user-base and datasets, modern Information Retrieval techniques are striving to become contextually-aware in order to provide users with a more satisfactory search experience. While text-only retrieval methods are significantly more accurate and faster to render results than purely visual retrieval methods, these latter provide a rich complementary medium which can be used to obtain relevant and different results from those obtained using text-only retrieval. Moreover, the visual retrieval methods can be used to learn the user's context and preferences, in particular the user's relevance feedback, and exploit them to narrow down the search to more accurate results. Despite the overall deficiency in precision of visual retrieval result, the top results are accurate enough to be used for query expansion, when expanded in a controlled manner.

The method we propose overcomes the usual pitfalls of visual retrieval:

1. The hardware barrier giving rise to prohibitively slow systems.

2. Results dominated by noise.

3. A significant gap between the low-level features and the semantics of the query.

In our thesis, the first barrier is overcome by employing a simple block-based visual features which outperforms a method based on MPEG-7 features specially at early precision (precision of the top results). For the second obstacle, lists from words semantically weighted according to their degree of relation to

the original query or to relevance feedback from example images are formed. These lists provide filters through which the confidence in the candidate results is assessed for inclusion in the results. This allows for more reliable Pseudo-Relevance Feedback (PRF). This technique is then used to bridge the third barrier; the semantic gap. It consists of a second step query, re-querying the data set with an query expanded with weighted words obtained from the initial query, and semantically filtered (SF) without human intervention.

We developed our PRF-SF method on the IAPR TC-12 benchmark dataset of 20,000 tourist images, obtaining promising results, and tested it on the different and much larger Belga benchmark dataset of approximately 500,000 news images originating from a different source. Our experiments confirmed the potential of the method in improving the overall Mean Average Precision, recall, as well as the level of diversity of the results measured using cluster recall.

# Acknowledgments

I would like to thank my advisors, Dr. Leila Kosseim and Dr. Sabine Bergler, for their continued moral, intellectual, and material support throughout the program. I am also grateful to the advisory committee for providing guidance and vetting of the research. Last but not least, I would like to thank my friends and colleagues at the Computational Linguistics at Concordia (CLaC) lab, and at Cold Spring Harbor laboratory.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

With the ever-increasing availability of digital images, image retrieval has become a common activity for professional, leisure and personal purposes. Journalists search for photos relating to a current news event in a press agency's photograph library, artists search their portfolio of images including both photographic and non-photographic images such as sketches and abstract art work, and computer and digital camera users often keep personal photographic collections that they need to search.

As these situations illustrate, image retrieval deals with a variety of image collections that can vary greatly with respect to size and content. A personal photographic collection typically contains a few thousand images, while the Associated Press Photograph Library contains more than 15 million images.[1] By comparison, the Internet has an estimated tens to hundreds of billions of images, many of which are stored in photo-sharing sites such as flickr (www.flickr.com) and Picasa (www.picasaweb.google.com), or social networks such as Facebook (www.facebook.com).

Today, most Image Retrieval engines rely solely on textual data and the associated annotations such as title, legend and comments, or metadata produced electronically, regardless of whether the user is looking for a text document, images or videos. This is typically true for image and video search. While this approach is efficient in terms of the time and resources required, it suffers important disadvantages due to the lack of metadata and the high costs associated with its creation by human annotators, as well as the inherent

[1] http://www.apimages.com/

ambiguity of cross-media searches.

Another important aspect of image retrieval is the domain of the collection. This can range from a completely closed domain of images that are very similar visually and semantically, such as mouse brain images, to open domain ad-hoc images of non-specific nature or topic, such as those on photo-sharing websites. Retrieval in closed-domain image collections can make use of the domain knowledge available. For example, searching mouse brain images can incorporate knowledge about the brain model in order to retrieve images from a specific region. This methodology is not feasible in ad-hoc image retrieval where no specific knowledge is readily available. In one of the earliest comprehensive surveys of content-based retrieval [Smeulders *et al.*, 2000], the open domain, referred to as *broad domain*, is described as having "unlimited and unpredictable variability in its appearance even for the same semantic meaning" as opposed to the limited and predictable variability of the *narrow* (closed) domain. [Datta *et al.*, 2008] further refines the categories of the scope of the data by dividing them into Personal, Domain-Specific, Enterprise, Archive and Web scopes. The goal of this thesis is to investigate the use of both image and text features to present more satisfactory results for queries on image repository, in the context of open-domain ad-hoc photographic collections. To translate the user's satisfaction into quantifiable measure, we use both traditional metrics of precision and recall, as well as the more recent metric *cluster recall* which measures the diversity of results in a presumed interactive retrieval process.

Studies on user behavior have found that users performing a search task tend to supply very short queries consisting of only a few terms [Goodrum and Spink, 1999]. This short text query, having already gone through one level of interpretation by the searcher of her own information need, and implicitly relying on her familiarity with the subject and linguistic vocabulary skills, undergoes more automatic levels of interpretation by the search engine. This process of multiple interpretations and disambiguation of the query often introduces compounded errors.

In image retrieval, the search can be done strictly using textual clues, visual features or a combination of both. The precision of strictly visual algorithms deteriorates rapidly. Moreover, due to the semantic gap between the low-level features and the higher-level concepts in the image (see Section 1.2), visual similarity

does not necessarily imply conceptual similarity. Even visual concept detection, still only practically viable for very general scenic criteria such as outdoors/indoors, day light/night etc. [Deselaers and Hanbury, 2009], falls short of capturing such complex contextual information as actions and scenarios, expressive clues such as feelings, and extravisual pragmatic characteristics such as motive. Content-based methods also require accurate representative example images and immense training corpora.

On the other hand, textual retrieval outperforms visual methods in speed and average precision but suffers important drawbacks, including the need for extensive annotation, and ambiguity at the interpretation level due to the image polysemy (the semantic ambiguity in the image [Heesch and Rüger, 2008]).

Despite the recent influx of research on image retrieval, a quality breakthrough is not yet in sight. The 2005 issue of ImageCLEF, the image retrieval benchmark of CLEF (Cross Language Evaluation Forum), reports a 41.35% highest Mean-Average Precision (MAP) for monolingual image queries in the ad-hoc retrieval task [Hoi *et al.*, 2005], while content-based visual retrieval achieves a meager 8% MAP [Chang *et al.*, 2005].

Research frequently cites the *Semantic gap*, the distance between low-level image representation and its semantics as the challenging source of complexity in content-based image retrieval. The semantic connotation of the search query is yet another factor requiring disambiguation. While earlier research focused mostly on either text-based or content-based retrieval, there has been a heightened interest in recent work in combining image and textual features as well as utilizing user-interactive techniques such as *Relevance feedback* in attempt to bridge this gap.

The remainder of this chapter motivates our proposed solution by introducing specific differences between image and text retrieval in relation to the *Semantic gap* then the various contextual aspects of Image Retrieval.

## 1.1 Image Retrieval

Image retrieval can be generally defined as a relevance ranking function f(q,D) that returns the list of most relevant ranked images (R) with respect to a query (q) from an image collection (D). The major divergence of image retrieval from textual information retrieval is that the documents that constitute the expected result list are images rather than documents or paragraphs of text. Moreover, in the case of image retrieval, queries

can belong to either modality. They can be textual metadata or examples of images to search for. Features such as dimensions, orientation, scale and rotation are also specific to image collections.

The type of the initial query, visual, textual or combined, as well as alternating between these types is also another source of information on the user's needs. For example when the user switches from text to visual query, it might indicate that she found particularly interesting results or that she is moving closer to her goal. On the contrary, switching from visual to text query might mean an inability to visually formulate the requirements of the search.

The dataset plays yet another role in determining the next appropriate step in an interactive search. Availability of associated text varies greatly from annotated images to text loosely floating in the vicinity of the image as in the case of web pages. Metadata and semi-structured data help greatly when available. In their absence, visual cues need to be incorporated. The size of the dataset and its degree of homogeneity influence the search process. For example, clustering the results might not make sense in very homogeneous sets and in small result sets where it would probably be more beneficial for the user to browse through the individual images. Section 4.1.7 uses two datasets to investigate the effects of the size and nature of the corpus on the retrieved results.

## 1.2 The Semantic Gap

As mentioned in Section 1, the *Semantic gap* is currently the main obstacle in Image Retrieval. The semantic gap has been traditionally defined as the distance between the semantic content of an image and its low-level representation. It can be viewed as an ambiguity in the sense of the image and the query. The challenge of information retrieval is to disambiguate the sense of a query with respect to a document collection. This is also true in the case of text retrieval. Table 1 draws a parallel between text and image ambiguity.

In text retrieval, at the lexical level, ambiguity can result from language identification. For example, the word "barn" in English means *"large building for storing farm products"* while in Norwegian it means *"a child"* [2]. This is why it is essential for a search engine accessing the whole WWW to provide a language

[2] http://translate.google.com/

Table 1: Parallel between Text and Image Ambiguity

| **Level** | **Text Features** | **Image Features** |
|---|---|---|
| **Lexical** | Language | Color, Texture |
| **Syntactic** | Bag of Words | Shape, Layout |
| **Semantic** | Word Sense Disambiguation | Image Sense Disambiguation |
| **Pragmatic** | Relevance Feedback | Relevance Feedback |

specification mechanism on the querying interface. In image retrieval, a corresponding ambiguity occurs on the level of color and texture. For example a search by color on a blue sky could return a blue sea instead. However adding the texture could retrieve the right images.

At the syntactic level, text search engines normally use a *bag of words* approach, an unordered list of the search terms, where the syntax does not affect the outcome of the search query (also *stop words* i.e. grammatical words not contributing to the semantics of the search are dropped). However, in some cases it is crucial to keep both the order and the exact phrasing of the search as in the case of searching for a book by name. Web search engines usually tackle this problem by providing some grouping operator like quotation marks around the phrase or dots between the search terms. In image retrieval, the syntax can be compared to the shape and layout structures. Knowing the exact object shapes or general layout of an image can help in its retrieval. However, this is not often the case, especially in the case of abstract concepts (e.g. war photos). This can be likened to knowing the exact phrase in text retrieval.

In previous work, content-based image retrieval has mostly relied on the above-mentioned levels of disambiguation, namely, color, texture, shape and layout. Nevertheless, ambiguity quite often occurs at the next levels, the semantic and pragmatic levels. In text retrieval, disambiguating at these levels involves using Word Sense Disambiguation (WSD) techniques. Specifically, synonyms and hypernyms (words more generic or broader in meaning) are used to expand the query and to ensure the retrieved documents are relevant to the required sense. By analogy, Image Sense Disambiguation (ISD) could be used to identify the semantic meaning the user is interested in. This method was used in [Bartolini, 2005] to contextualize image queries by presenting the user with results and requerying the image repository with the relevant ones. For example, a query on *Tank* can return water tanks or war tanks, by checking the relevant images and requerying the database only tanks of the desired type are retrieved. Although Voorhees found that the Is-A relation from

WordNet [Fellbaum, 1998] resulted in deteriorating the results since queries are often too short to provide appropriate context for disambiguation [Voorhees, 1993].

At the pragmatic level, *relevance feedback* techniques can be used to respond to specific user requirements. This is important since searching is often an evolving activity. Users might have only a vague idea of their search target in the beginning and progressively discover it through browsing and inspection of the search results.

The next section sheds light on the search process itself from the user perspective.

## 1.3 Image Retrieval Stages

From the user's point of view, image retrieval follows a pipeline which can generally involve some or all of the following steps: a query formulation stage, a results presentation stage, a browsing stage and a relevance feedback stage. This section introduces these stages.

### 1.3.1 Query Formulation

Query formulation and reformulation allow the user to express their need from the dataset. This could use both text-based and content-based methods. The ability to formulate queries depends on the user's experience and knowledge as well as the nature of the data and the task. More advanced users might tend, for example, to know shortcuts to easily express their needs. Some search engines provide an Advanced Search option which leads to an interface with more options, such as retrieving images in a specific size range or from specific sites. Also multiple criteria involving the use of logical operators "and", "or" and "not" can be applied through the advanced options. Queries in image retrieval can vary greatly in their level of difficulty. The 2010 ImageCLEF benchmark [Agosti *et al.*, 2010] (discussed in Section 3.1) divides the difficulty of topics into four levels according to the Mean Average Precision (MAP) (discussed in Section 3.4.1) of the results of a given query, with the MAP ranging from under 0.1 to over 0.3. Easier topics tend to include more named-entities which can benefit from text-based retrieval techniques, while harder topics contain semantic and visual cues [Agosti *et al.*, 2010].

### 1.3.2 Presentation

The presentation of the data affects the effectiveness of the search. Search engines presenting the results as one ordered list of thumbnailed images often make the wrong assumption that the user will find the desired image at the top of the result set. When this is not the case, it could be nearly impossible to succeed in the search if the result set contains thousands of images. Deciding on the relevance of the results of an image search takes significantly less time than these of a text search.

### 1.3.3 Browsing

Browsing is an interactive step in image retrieval. It allows users to go through the retrieved images and possibly provide relevance feedback. Browsing can use expandable thumbnails, hierarchical menus of concepts as in [Clough *et al.*, 2005a] and [Petrelli and Clough, 2005].

### 1.3.4 Relevance Feedback

Relevance feedback is the process whereby the retrieval engine receives feedback about the relevance of the initial results returned, and attempts to incorporate this information to produce better results by performing another search. This process can be repeated iteratively as many times as desired until images adequate to the user's needs are found or no further improvement is achieved. Relevance feedback can be done either manually, involving input from the user, or automatically without human intervention. Three types of relevance feedback are typically used: User relevance feedback, Pseudo-relevance feedback, and Semantic relatedness.

#### User Relevance Feedback

User relevance feedback is a possible interactive step in information retrieval where the user is given the opportunity to indicate the relevance of the retrieved documents and possibly its degree of relevance. In the case of boolean relevance feedback, the user can simply assign a boolean relevance judgment to a given image in the result set. The input of the user could also be qualitative (e.g. very relevant/relevant/not so relevant/not relevant) or quantitative by assigning a score to the returned images. The user might also be able to indicate

both positive and negative relevance.

#### Pseudo-Relevance Feedback

Pseudo-relevance feedback, also referred to as blind or auto-relevance feedback, is the process of automatically re-querying the dataset without intervention from the user. In this situation, the results returned from an initial run (most likely the best results) are used to extract more information either visual or textual if available and a new query is sent to the database. Pseudo-relevance feedback has proven effective in text IR as well as Image IR [Chang *et al.*, 2005].

#### Semantic Relatedness

Queries, expressed in natural language, could possibly have numerous semantic equivalents. The same applies to the document collection. In order to deal with these semantic variations, several techniques have been proposed. On the querying end, query expansion techniques can be used to augment the query with relevant terms such as synonyms (e.g. [Voorhees, 1994]). This approach, however, needs careful word-sense disambiguation techniques in order to avoid introducing noisy terms that lower the precision. Synonyms, often, cannot be used interchangeably.

The method proposed by this dissertation for enhancing the precision and recall of photographic image retrieval involves the use of both pseudo-relevance feedback and semantic relatedness to improve the precision of the retrieved results as well as the number of relevant results retrieved. This will be described in details in Chapter 5.

## 1.4 Thesis Motivation

This dissertation aims to improve the quality of image retrieval results. Our hypothesis is that incorporating visual and textual features in the search through pseudo-relevance feedback can significantly improve the results over single-modality search, provided adequate semantic filtering is employed. The methodology we followed is carrying out experiments employing single-modality, and comparing the results of these to fusion

approaches using standard benchmarks. We measure the improvement in quality using standard IR measures such as Mean Average Precision (MAP), Recall, as well as the precision of the highest ranked results of image retrieval (See Chapter 3). In order to achieve this improvement, as shown in Figure 1, the retrieval process is studied from five angles. Since these angles are cumulative, they are evaluated at each stage of the retrieval using a well-studied benchmark: the ImageCLEFPhoto dataset for the three years from 2007 to 2009.

1. The first angle that we studied is the text-based retrieval. Despite its short comings discussed in Section 2.1, text-based retrieval is still considered the cornerstone of the retrieval process due to the far better results that could be achieved using text-only retrieval compared to pure visual retrieval besides the possibility of the lack of an example image. Under the realistic conditions of the existence of sparse annotations, the vector-space model (section 2.1.2) and the probabilistic model (section 2.1.3) are compared. A simple clustering of the collection to augment the results is also explored.

2. The second angle investigated is the pure content-based retrieval, relying solely on visual features. Research in thes area is commonly referred to as Content-Based Image Retrieval (CBIR), since it relies mostly on the content of the images themselves. As in the case of text-based retrieval, two approaches are compared: one relying on more sophisticated features based on MPEG-7 descriptors [Martínez, 2004] [Lux and Granitzer, 2005] (discussed in Section 4.2.1), while the other uses simple block-based statistical features.

3. The third angle constitutes the fusion of the results from text-based and content-based retrieval. According to [Clinchant *et al.*, 2010] using appropriate fusion methods is a precondition to improving the precision of the retrieved results; for this reason we considered this angle important to investigate.

4. The fourth angle of this research which is the query expansion using semantic filtering of the results of the fusion as an attempt to narrow the semantic gap between the low-level visual features and the semantic content of the image. In this respect, different semantic relations are examined for their potential in improving the results.

5. The fifth and final angle of this thesis deals with a specific fusion technique which is auto-relevance feedback. Auto-relevance feedback is a step performed entirely without the user's intervention where the system attempts to evaluate the relevance of the retrieved results, and take actions to improve the user's query accordingly.

To assess the effect of each angle, all five angles have been evaluated using the standard collections that were employed in the ImageCLEFPhoto task of the CLEF benchmark in 2007 [Grubinger *et al.*, 2007], as well as 2008 [Arni *et al.*, 2008 printed in 2009] and 2009 [Peters *et al.*, 2010].

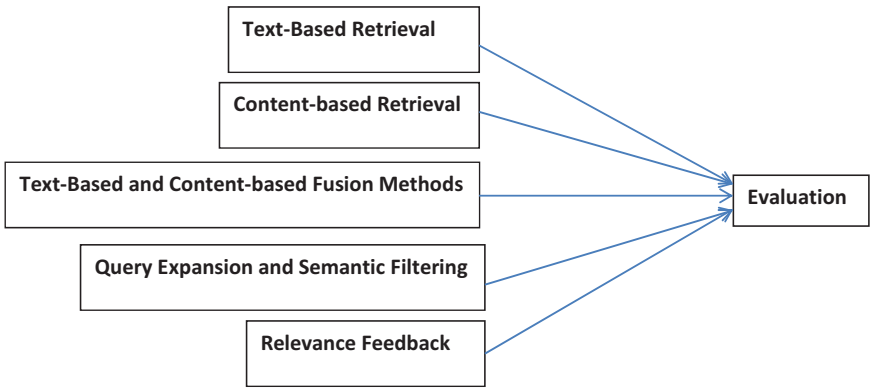


Figure 1: Structure of the Thesis: the Five Angles Studied in this Thesis

## 1.5 Problems with Text-Only Image Retrieval

Due to the prevalence of text search engines on the Internet, the ability to express a textual query is more developed in most users than that of formulating an exact visual query. However, using only text to specify a user's information need inhibits the ability to benefit from the richness of the visual medium, and can cause

several problems in the results.

Moreover, according to recent research trends in interactive image retrieval, it is considered beneficial to diversify the results over the set of possible different interpretations of a given query. Providing the searcher with an initial result set consisting of documents belonging to as many possible interpretations of a query's meaning, could provide the means to bridge the semantic gap. For example, searching on the term **Fringe** could imply that the user is looking for a specific sense of fringe (e.g. fringe benefits, fringe art) rather than the general sense **Edge**. An effective search engine therefore need look not only for the term **Fringe** using its more widely used meaning, but also occurrences of the various other senses.

The relationship between ambiguity and diversification of the results has been addressed in several studies, such as [Agrawal *et al.*, 2009] and [van Zwol *et al.*, 2008]. To illustrate, consider the problem of expressing a search on a large white house that is not the White House. The first page of results of the search query example on **white house** submitted to the Google search engine is shown in Figure 2 while Figures 3 and 4 show pages 23 and 49 of the same results. Indeed the same results! Despite returning 794,000 results for white house and 412,000 for the two words joined, Google will only display 50 pages of results. The consequence is that the 50 pages are dominated by White House-related images with very few and in-between other white houses, which does not respond to the users query. Even a more experienced user posing a better-formed specific query with "-Washington" directive to try to exclude White House images will not get the desired result with 350,000 images still mostly dominated by the White House.

Another problem with text-only image retrieval is the lack of a mechanism to extract and account for the visual information provided by the user in their relevance feedback.

An alternative approach is to retrieve results based on a combination of visual and text features. This could lead to mixed results with mostly the White House images, and other images of white houses and possibly a few more with people related to the White House, images of reports from the White House etc... The user can then browse through a smaller summary of the results presented by representative images of each cluster. These representative images can be from the center but also the extremes of the cluster to give the user a chance to objectively evaluate their relevance.

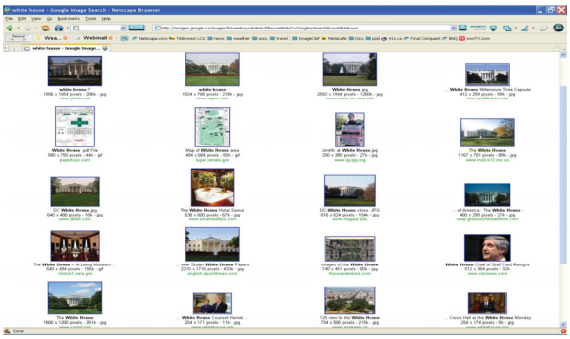

Figure 2: First page results from Google Image Search on “White House”

## 1.6 Contributions

The main outcomes of the research presented in this thesis are the following:

- Enhancing the precision, recall and diversity of image retrieval using inter-media fusion. This outcome is the result of studying all five angles in Figure 1 combined, and can be deduced from the evaluation presented in Chapter 5, with a detailed insight into how the approach achieves this improvement using examples. The method used achieves higher precision than any of the results reported in the ImageCLEF 2008 and 2009 campaigns, which employed two different datasets.

- Proposing a robust method for inter-media query expansion that functions on different datasets. An outcome of the fourth angle of research according to Figure 1

- Investigating and successfully incorporating semantic expansion and semantic filtering of text queries. Another outcome of the fourth explored angle of research, this outcome demonstrates the feasibility of introducing query expansions that that are not too noisy to be effective.

- Promoting diversity in image retrieval results by incorporating both text and visual features. This represents the outcome of the third angle of research and is covered in Chapter 5.

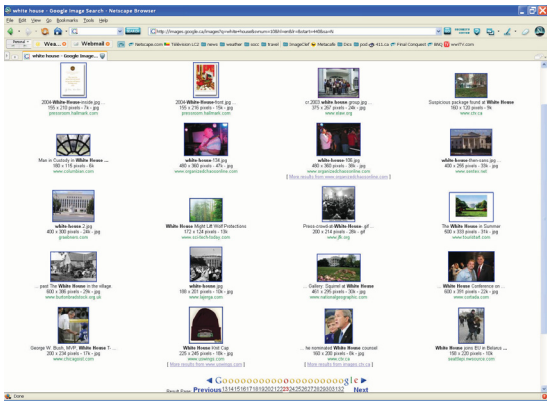

Figure 3: Page 23 results from Google Image Search on “White House”

- Comparing different text retrieval models within the context of image retrieval task. This corresponds to the first angle of Figure 1 and is covered in Section 4.1

- Proposing a simple low-cost visual retrieval method that outperforms MPEG-7 descriptors. The outcome of the second angle of research is presented in Section 4.2

- Participating in three benchmarking campaigns for image retrieval (ImageCLEF), to formally evaluate the results in comparison to other methods.

## 1.7 Organization of the thesis

The rest of this thesis is divided as follows: Chapter two is a brief overview of previous related work in the field of Image Retrieval. Chapter three includes a description of the evaluation resources, the metrics, corpora and queries used in benchmarking our method. Chapter four details the single-medium retrieval underlying our method, in addition to comparing different models for text (probabilistic and vector-based) and visual retrieval (MPEG-7 descriptors and a block-based method). Chapter five describes the complete semantic inter-media fusion method using pseudo-Relevance Feedback and Semantic Filtering (PRF-SF), and demonstrates some actual examples of applying the method, and Chapter six concludes the thesis pinpointing

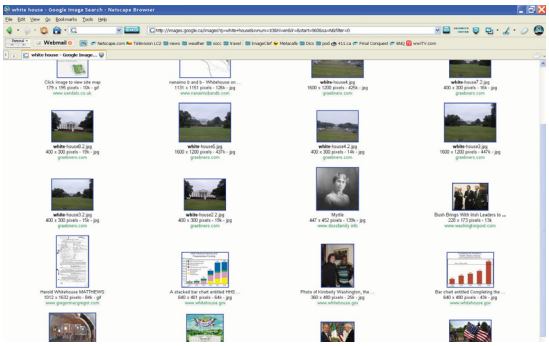

Figure 4: Page 49 results from Google Image Search on “White House”

the advantages of our approach.

# Chapter 2

# Related Work

As an extension to text retrieval, image retrieval dates back to the 1970s. Attempts to introduce content-based methodologies started in the 1990s with the advent of the Internet and large image collections [Rui and Huang, 1999]. This chapter introduces previous work covering different ways of approaching the five angles of research introduced in Chapter 1. First discussed is text-based image retrieval (angle 1), followed by content-based image retrieval (CBIR-angle 2) , then the different ways for the combination of these methods including relevance feedback are presented (angles 3,4, and 5). Finally, despite that experiments presented here rely on Pseudo-relevance feedback, the outcomes of the research in this thesis can be extended to user relevance-feedback framework. We therefore present research in contextual Image Retrieval in Section 2.5.

## 2.1 Text-based Image Retrieval

General search engines on the web like Google [Google, 2006], Yahoo [Yahoo, 2006], and MSN [MSN, 2006] rely mostly on text-based retrieval methods for image queries. The text surrounding the image and/or text in metadata tags is considered relevant to the image. Text-based retrieval is based on the same information retrieval (IR) techniques used for linguistic processing. Steps in text-based retrieval include removing stop words, tagging named-entities, stemming words and sometimes using synonyms, hypernyms and other semantic relations to improve the chance of finding relevant matches. The aim of textual information retrieval

is to satisfy a user's search need by ranking highest the most relevant results to the text query in the fewest number of steps.

Where available, text is invaluable in image retrieval, giving more accurate and faster results. However, this is rarely the case. The most significant obstacle in text-based image retrieval is the lack of annotations. To address this problem, [von Ahn and Dabbish, 2004] created an online interactive game that relies on the agreement of two players to gather image annotations that can then be used for learning[1]. The other challenge in text-based retrieval is that even in the presence of annotations, the weighting schemes used in text IR cannot be applied seamlessly to image retrieval. For example, the concept of term frequency (TF) is not applicable to annotations unless extracted from an actual text document containing many more words than annotations. The next sections will describe text information retrieval, as it is the basis of most work in image retrieval. The most popular classical text information retrieval models can be categorized into boolean, vector-based and probabilistic models.

### 2.1.1 The Boolean Model

The simplest of IR models is the Boolean model which represents the query as a boolean expression and assigns a boolean value to the terms in the collection indicating their presence or absence in a document. Similarly, relevance judgment of the documents in the collection is the boolean outcome of the query expression. As such, boolean models are not adequately capable of describing the extent to which a document is pertinent relative to a given query, which impedes the ranking of results. In image retrieval, boolean querying was used in [Fauqueur and Boujemaa, 2004] to help the searcher compose a mental image of her information need using image regions. While suitable for database querying, the boolean model is inadequate for real-life image retrieval applications, and hence, it will not discussed further in this thesis.

### 2.1.2 The Vector Space Model

In vector space models, documents and queries are represented as a vector of features. The vector space model was introduced in 1975 in [Salton *et al.*, 1975]. The most common vector space model is the TF-IDF, where

[1] http://www.gwap.com/gwap/gamesPreview/espgame/

the features are words weighted using term frequencies (TF) multiplied by the inverse document frequency (IDF). Term Frequency (TF) indicates the importance of a term in a document by counting its occurrences. It is usually normalized by the length of the document. The Inverse Document Frequency (IDF) as a measure of the specificity of a term was first described in [Jones, 1972]. To compare the vectors, one of the frequently used measures is the cosine distance. Vector-space models are often the model of choice for representing bag of words representation. A disadvantage of this model is the invalid assumption of independence of the terms.

The TF-IDF model has been frequently used in the context of the ImageCLEF benchmark (described in Chapter 3). [Hoi *et al.*, 2005] found that using a language model based on relative entropy or Kullback-Leibler divergence [Kullback and Leibler, 1951] achieved better results than TF-IDF on the ImageCLEF data. However, this might well be the result of the size and the nature of the dataset which is limited in nature and belongs to the same domain. [Fakeri-Tabrizi *et al.*, 2010] also found that a probabilistic language-based model (see Section 2.1.3) gave better early precision at 5 and 10 documents retrieved than the TF-IDF model.

### 2.1.3 Probabilistic Models

The ranking functions of probabilistic information retrieval models assign a probability to a document's relevance to a query based on uncertainty. Work on probabilistic models began in the 1970s and 1980s such as in [Robertson and Sparck Jones, 1976]. Probabilistic models work with two independence assumptions: that the relevance of a document is independent of the relevance of another, and that probabilities of terms in a document are conditionally independent [Manning *et al.*, 2008]. One of the most successful probabilistic ranking functions is Okapi BM25 [Robertson *et al.*, 1996]. A recent review of the Probabilistic Relevance Framework can be found in [Robertson and Zaragoza, 2009].

A probabilistic model based on language models for the text retrieval component of an image retrieval system is introduced in [Westerveld *et al.*, 2003]. The model gave mixed results when tested on an easier Corel dataset [Westerveld and de Vries, 2003a] and the harder TRECVID data [Westerveld and de Vries, 2003b] . According to the authors, this was due to Corel's dataset being not realistic and much easier for the model.

### 2.1.4 Ontologies

Ontologies are a form of knowledge representation that models concepts and the relationships between them [Gruber, 1993]. There have been several attempts to incorporate ontologies in the context of image retrieval. [Magesh and Thangaraj, 2011] created a general ontology hierarchy and tested it on an image collection of 2000 images. They reported an improvement in the results. RetrievOnto [Popescu *et al.*, 2007] used the WordNet [Fellbaum, 1998] hierarchy starting from the term *placental* to structure the dataset, improving the results from CBIR. Ontologies can be used in the retrieval process to overcome the problem of polysemy, or for query expansion to add relevant terms to the query. An example can be found in [Cumbreras *et al.*, 2009] who successfully used the MeSH ontology and the UMLS thesaurus to perform query expansion.

### 2.1.5 Text Query Expansion

Text query expansion refers to adding words to a text query to increase the likelihood of finding relevant documents (recall) and accuracy (precision) of retrieval. As discussed in Chapter 1, semantic analysis of a query, including word sense disambiguation and adding synonyms can lead to undesirable results by adding noise to a given query. [Clinchant *et al.*, 2010] used a textual entailment probabilistic model to expand the query with the terms most related to the given query terms. [Martínez-Fernández *et al.*, 2005] experimented unsuccessfully with the ImageCLEF data using hypernyms to expand a query. Query expansion remains an open research topic in the context of image retrieval, which is addressed in this dissertation.

### 2.1.6 Tools

There are many open-source tools available for text-based processing. These include full-fledged search engines such as Apache Lucene [Hatcher and Gospodnetic, 2004], Terrier [Ounis *et al.*, 2006], Sphinx[2], and Xapian[3], clustering engines such as Carrot2 [Osinski and Weiss, 2005], syntactic parsers such as the Stanford Parser [de Marneffe *et al.*, 2006], part of speech taggers such as the Brill tagger [Brill, 1992], and complete

[2] http://sphinxsearch.com/
[3] http://xapian.org/

integrated toolkits such the NLTK[4] [Bird *et al.*, 2009]. Other tools freely available include named-entity taggers such as the Illinois Named Entity Tagger [Ratinov and Roth, 2009] [5], Gazetteers which include the names of places and people, dictionaries, thesauri and lexical databases such as WordNet [Fellbaum, 1998]. GATE is a framework which includes some NLP tools and allows the incorporation of others[6]. As we will see in Chapter 4, we have taken advantage of Lucene, Terrier, and WordNet to experiment with our work.

## 2.2 Content-Based Image Retrieval

As opposed to text-based retrieval (Section 2.1), Content-Based Image Retrieval (CBIR) refers to using the low-level features of images for identifying the ones relevant to a given query. The query can be either text or image, also known as Query By Example (QBE). A key survey of content-based work in the years 1990-2000 appears in [Smeulders *et al.*, 2000]. More recent comprehensive surveys of the issues and trends in CBIR can be found in [Lew *et al.*, 2006] and [Datta *et al.*, 2008]. There are three main components to a CBIR system:

- Feature Extraction
- Feature Representation
- Similarity Modeling

This section reviews these components and their impact on the retrieval process.

### 2.2.1 Feature Extraction

Feature extraction, also often referred to as *Visual Signature Extraction*, is the process of capturing visual characteristics or features of the image.

#### Global vs. Local Features

Global features are those extracted from the whole image, while local features are extracted from specific regions. [Douze *et al.*, 2009] evaluate the use of the global descriptors first proposed in [Oliva and Torralba, 2001],

[4] http://www.nltk.org/
[5] http://cogcomp.cs.illinois.edu/page/software_view/4
[6] http://gate.ac.uk/

on web image retrieval. Global descriptors do not require image segmentation, and thus are relatively less resource intensive.

In general, the basic features most often used in content-based retrieval can be grouped under color, texture and shape features [Goodrum, 2000].

### Color Features

Color features are the simplest and most frequently used feature in content-based retrieval [Squire *et al.*, 1998] and can be considered the baseline in content-based retrieval. Colors have the advantage of being invariant to rotation and scaling. Color features are also the least computationally intensive of the features. However, global color features often lead to inaccurate results, so region-based color features are sometimes needed. Color distribution is often represented as histograms of the colors in a specific color space. A color space represents colors as tuples of values with a mapping function to an absolute reference color model. [Wang *et al.*, 1997] employed hierarchical and k-means clustering techniques to colors to improve the efficiency of retrieval. [Mandal *et al.*, 1996] and [Stricker and Orengo, 1995] successfully used the first moments of color histograms as features rather than the histograms themselves.

### Texture Features

A texture is a repeated pattern. Texture features can be captured through a variety of visual qualities like coarseness, directionality, roughness and contrast. [Deselaers *et al.*, 2004] experimented with the set of features known as Tamura features described first in [Tamura *et al.*, 1978] and *coarseness*, *contrast* and *directionality* were the most significant in describing texture.

### Shape Features

Edge, curve and corner detection are used to represent shapes in images. Due to the complexity of different possible combinations, these features are often only successful in closed-domain problems.

### 2.2.2 Feature Representation

While *Feature Extraction* deals with capturing the most salient features to represent an image, the representation of the feature plays a crucial part in modeling the visual retrieval process.

#### The Bag of Features Approach

The Bag of Features approach in image retrieval draws its inspiration from the equivalent Bag of Words text retrieval model. In this analogy, the image is the equivalent of the document, while specific areas of the image constitute the visual words. The frequency of words constitute the features themselves. One of the first examples of using this approach is described in [Sivic and Zisserman, 2003]. Recently, [Douze *et al.*, 2011] improved on this approach by combining it with Fisher Vectors.

One important disadvantage of the bag of words model is that it does not account for word order. Thus, the sentences "my other results can be considered significant" and "results can be considered my significant other" are equivalent. This illustrates that the bag of words model can not allow for techniques utilizing collocations in finding the most relevant documents. Similarly in images, the bag of features approach does not account for the spatial distribution of the image patches, thus losing important semantic information.

#### Spatial Features

Contrary to the bag of features model, models employing spatial features contain information about the location of a feature. This can be local information similar to the n-gram word models used in text retrieval, or global features which model the spatial location of the feature in the whole image. The MPEG-7 standard (discussed in Section 4.2.1) includes both types of features. For example the color structure descriptor is a histogram of a moving window of 8x8 pixels which captures the local structure of the color distribution.

### 2.2.3 Similarity Measures

Feature extraction and representation are the first two steps in a content-based image retrieval system. The third is modeling the similarity between the query and images in the collection. According to [Datta *et al.*, 2008],

the concept of similarity in image retrieval can belong to one of the following categories:

- Vector/non-vector/ensemble representations
- Local features/global features/combination
- Linear space/Non-linear manifold
- Stochastic/Fuzzy/Deterministic
- Supervised/Semi-Supervised/Unsupervised

### 2.2.4 Systems Employing Content-based Retrieval

Several systems employing content-based retrieval have been developed and deployed on the web. For example, QBIC (Query By Image Content) [Flickner *et al.*, 1997] was previously available commercially from IBM. Webseek described in [Smith and Chang, 1997] is a content-based web image and video search engine[7]. Viper/Gift is an open-source image finding tool developed at the University of Geneva [Squire *et al.*, 1999] and available under GNU License Terms[8]. Other systems include pictoseek [Gevers and Smeulders, 2000], Mars [Rui *et al.*, 1997] which uses relevance feedback, Cortina [Quack *et al.*, 2004], and LIRE which uses MPEG-7 descriptors [Lux and Granitzer, 2005] and which was employed in some of the experiments in this thesis.

## 2.3 Text-based and Content-based Retrieval Combined

As discussed earlier, most text-based search engines rely on meta-data and other textual data associated with the image, obtaining faster and more accurate results than content-based methods. Recent research has been exploring ways to incorporate content-based methods in text-based image retrieval [Wang *et al.*, 2008]. The fusion between visual and textual methods can happen early or late in the retrieval process. Early fusion is

[7] http://persia.ee.columbia.edu:8008/
[8] ftp://ftp.gnu.org/gnu/gift

applied before or during the actual retrieval steps, while late fusion incorporates the results obtained from both retrieval engines.

Different approaches have been explored to combine text-based and content-based retrieval. According to [Liu *et al.*, 2007] these methods can be categorized into five different approaches:

1. Defining high-level concepts using an object ontology
2. Associating low-level features with semantic concepts using machine learning
3. User relevance feedback
4. Generating semantic template to support high-level image retrieval
5. WWW fusion of text and visual content

In another approach, [Besançon and Millet, 2005] experimented with merging results from content-based and text-based systems using different weights. The merged results increased the Mean Average Precision (MAP) by 17%-18%. However, the weights were chosen empirically and reported to not fit different data from the previous year. According to the authors, it is not clear how to tune the merging strategy.

In [Cascia *et al.*, 1998], Latent Semantic Indexing (LSI) on HTML documents is combined with visual statistics namely color histogram and dominant orientation histogram as a global feature vector. A relevance feedback scheme is then employed to respond to Query By Examples. The results obtained suggest that combining text and image data significantly improve results over those from a single medium. Another research that used Latent Semantic Indexing [Zhao and Grosk, 2002] combined with color/anglogram histograms arrived at a similar conclusion, although using text queries in this case.

XRCE [Clinchant *et al.*, 2010] experimented with different fusion methods in the ImageCLEF 2010 Wikipedia ad-hoc retrieval task, a multi-lingual and multimedia retrieval task [Agosti *et al.*, 2010] (described in Section 3.3.4). They observed that selecting an appropriate fusion method is critical to improving the results over the text-only based retrieval, while the visual-only retrieval performs very poorly in comparison. The fusion methods they experimented with includes re-ranking the highest 2000 text-only retrieval results based on visual similarity. This approach resulted in an improvement of the precision of the first 20 retrieved results

(P20) while decreasing the mean average precision (MAP). Their most successful attempt was obtained by aggregating the results from this list with the text-only results using a weighted mean average. For the photo retrieval task, XRCE obtained the highest results in the benchmark using a system that employs query expansion relying on term co-occurrence measured by the Chi-Square statistic to denote term similarity and textual entailment.

### 2.3.1 Relevance Feedback

In addition to the approaches described above, pseudo-relevance feedback (see Section 1.3.4) has also been exploited in multi-modal retrieval. An example can be found in [Chang *et al.*, 2005] where the dataset is re-queried with annotations from the top two images retrieved from an initial visual run that used the Viper/Gift visual retrieval engine. This technique improved the results from 8% to 34.5% MAP. However, it relies on the existence of annotated data (the dataset used is ImageCLEF which is annotated for the most part).

One of the early systems to incorporate a Relevance Feedback mechanism in content-based image retrieval was Mars [Rui *et al.*, 1997], that compared TF-IDF to Gaussian Normalization to estimate the weights of the features in the feedback step. Experimenting on a dataset of 384 texture images, the authors found that TF-IDF performs better than using TF only, and that while TF-IDF generally outperforms Gaussian normalization, there are cases where normalization leads to better results.

[Maillot *et al.*, 2006] investigated pseudo-relevance feedback and fusion methods on the same dataset used in the experiments, the IAPR TC-12 collection. They reported a precision gain with feedback but not with fusion. In the final phase, the results from the text and visual queries are post-fused through a re-ranking mechanism to increase recall and ensure diversity and coverage.

## 2.4 Clustering

According to [Jain *et al.*, 1999] *"There is no clustering technique that is universally applicable in uncovering the variety of structures present in multidimensional data sets"*. The motivation behind the use of clustering to present search results is to decrease the information load on the user and simplify the browsing process

leading to a more effective and faster search. Clustering can also be used in bulk annotation, by assigning the same annotations to a given cluster. The text information retrieval community has already experimented successfully with clustering techniques (for example [Leouski and Croft, 1996] and [Zamir and Etzioni, 1999]), and so it is worth exploring its use for image retrieval.

Clustering is the process of classifying patterns according to their inherent structural similarities. Clustering is an unsupervised learning method, meaning no prior statistical model is used in the process. When some information is available, the process is often referred to as semi-supervised clustering. There is a long tradition of using clustering analysis in various fields such as psychology, biology and marketing. Pattern recognition, image segmentation and machine learning are some of the areas that make use of clustering techniques. Clustering involves the following parameters:

- Data Representation: The scale and types used to represent the data to the clustering algorithm. Data can be represented on qualitative or quantitative scales. Quantitative data can be discrete, continuous or interval and qualitative data can be ordinal or nominal.
- Feature Selection: The specific features that will be used to differentiate the different classes of data.
- Feature Extraction: Feature extraction involves further processing of the selected features. This could be needed to arrive at a better or more efficient data representation.
- Proximity/Similarity Measure: It is essential in clustering to measure the distance between patterns using a proximity measure. Common measures include the Euclidean distance and the Mahalanobis distance.
- Data Abstraction: After the data has been clustered it is often useful to use an abstraction of the data for presentation convenience or further processing.
- Clustering Evaluation: In some cases there is not necessarily one best clustering, and the judgment of the quality of clustering could be subjective.

Clustering techniques can be categorized into hierarchical and partitional algorithms. Hierarchical algorithms output nested hierarchies of clusters, while partitional ones produce only one partition scheme.

Hierarchical clustering algorithms are further classified into agglomerative and divisive ones. Agglomerative algorithms start with each object in its own cluster and proceeds by grouping objects and reducing the number of clusters. Conversely, divisive algorithms start with all objects in one cluster and attempts to break the clusters down. Hierarchical algorithms can also be divided into single link or complete link. In single link, the distance between clusters is the minimum of distances between all pairs drawn from two clusters, while in complete link the distance between two clusters is the maximum between all pairs of patterns.

Partitional algorithms are divided into squared error, graph theoretic, mixture resolving and mode seeking methods. The squared error method uses an initial partition configuration and a criterion function at which to stop attempting to readjust clusters. Due to its simplicity and low time complexity, k-Means is one of the most popular squared error clustering algorithms, where k cluster centers are chosen at the beginning. Graph theoretic clustering algorithms attempt to construct the Minimum Spanning Tree (MST) of the data and deleting the longest edges to obtain the clusters. Mixture-resolving and mode-seeking algorithms are density-based approaches using statistical parametric methods such as Expectation Maximization (EM) or non-parametric methods such as multidimensional histograms.

Data abstraction involves generating cluster descriptions. These can be representative points, such as the centroids or furthest points in the cluster, by nodes in a classification tree or conjunctive logical expressions. A detailed description of clustering techniques can be found in [Jain and Dubes, 1988] and a comparative study in [Jain *et al.*, 1999].

In image retrieval, clustering has been applied in various ways. [Yin *et al.*, 2003], for example, grouped jointly-labeled images using relevance feedback from users of a search system, with the purpose of using the clustering for annotation. [Inkpen *et al.*, 2009] experimented with k-means clustering, hierarchical agglomerative clustering, and a novel method based on searching the WordNet relationns *Hypernymy* and *is an instance of* to form clusters of the results. This method outperformed the classical clustering algorithms, especially for images involving geographic locations.

Other examples can be found in [Sunayama *et al.*, 2004], who clustered images using labels from the surrounding HTML text. [Chen *et al.*, 2003] applied clustering to content-based image retrieval using the

Normalized Cut (NCut) algorithm under a graph representation. Another spectral clustering algorithm *Locality Preserving Clustering (LPC)* was introduced in [Zheng *et al.*, 2004] and found to be more efficient than NCut for image data.

There is very little material in the literature on clustering using both content-based and text-based features. [Cai *et al.*, 2004b] and [Cai *et al.*, 2004a] describe successive clustering applied on text features then image features. The textual features comprised a vision based text segment as well as the link information while the *Color Texture Moments (CTM)*, a combined representation of color and texture were chosen for visual features. Other research combining simultaneously image and textual features includes [Li *et al.*, 2005] and [Gao *et al.*, 2005] from Microsoft Research Asia, both using co-clustering techniques.

## 2.5 Context-based Retrieval

Context-based retrieval refers to the incorporation of contextual elements in the search process. In order to overcome the obstacle of understanding the query, some search engines experimented with building user models, based on roles and professional interests. However no significant improvement was observed [Goren-Bar *et al.*, 2001]. For this reason, researchers have started experimenting with context-based information retrieval. More specifically, context refers to information which can not be deduced from the query, but forms part of the environment of the query, such as the goal of the query, and in a broad sense it encompasses the user who makes the query. IRiX [Ingwersen and Järvelin, 2005], the information retrieval in context workshop held as part of SIGIR (Special Interest Group on Information Retrieval) attempted to explore the central themes of contextual information retrieval. Several studies have been conducted on users' needs in image retrieval. Some of these were targeted to specific groups like art directors [Garber and Grunes, 1992], journalists [Markkula and Sormunen, 2000] and graphic designers [Rodden *et al.*, 2001], while others involved general purpose image searching as in [Armitage and Enser, 1997]. Models of information behavior, as those described in [Pharo, 2004] and [Pharo and Järvelin, 2006] promise to adapt to the different and evolving contextual factors of the search process.

The context in image search varies greatly depending on the user, task, query and dataset. The combination of these factors can lead to a very different "ideal" search process from the point of view of the user. Users from diverse background categories tend to have special expectations. For example, art directors might find the "artistic concept" of the image of considerable significance [Garber and Grunes, 1992] and spend substantial effort choosing the "best" image among many alternatives. Journalists on the other hand may be satisfied with "acceptable" selections [Markkula and Sormunen, 2000]. [Rodden *et al.*, 2001] found that organizing thumbnails by visual similarity helps graphic designers in their real-life work tasks.

Recent directions in text information retrieval research show a shift in focus from content-based approaches through user modeling, and finally to context modeling. Precisely understanding what the user is searching for could improve performance. Another important aspect in search tasks is the knowledge or expectation level of the user. This can vary from absolute lack of knowledge to a very specific expectation of a previously-seen image (one of the scenarios used in the interactive task of one of the earliest benchmarks for image retrieval, ImageCLEF 2003 (see Section 3.1). Presenting diversified results, with different levels of relevance, could help the users in deciding on their next level in interactive search.

To the best of our knowledge, there is no research significantly relying on a specific search-behavior model for query refinement. This might be due to the costly resources required for user training and evaluation involved [Müller *et al.*, 2005b]. In fact, only two groups participated in the user-centered interactive retrieval task of ImageCLEF 2004 [Müller *et al.*, 2005a], with the same low participation continuing in 2005 [Gonzalo *et al.*, 2005]. This underlines the divide between theoretical and practical approaches in the field of image retrieval. While the first has been traditionally embraced by academic research groups (which is the case of most -if not all- groups participating in ImageCLEF ), industry research has favored user-targeted systems (e.g. [Li *et al.*, 2005]). An evidence of this is the use of WWW images in industry research (e.g. [Cai *et al.*, 2004b] and [Cai *et al.*, 2004a]) rather than specific collections as in the case of ImageCLEF so far. In addition, efficiency concerns, such as processing time, are a more significant factor for industry research (as in [Gao *et al.*, 2005]).

*Prism* [Leake and Scherle, 2001] is a search engine which attempts to extract contextual information using the Watson method [Budzik and Hammond, 2000] to monitor the user's activities in standard applications like word-processing. The Watson method makes use of style characteristics of words such as emphasized text, as well as the location of words in the document being authored as contextual indications of the importance of these words for queries. It then uses heuristics and traditional information retrieval methods (TF-IDF) (see Section 2.1.2) to infer the selection of specialized search engines to which it directs the user's query. Results from Prism suggest that using contextual information for this task can improve the retrieved results' usefulness.

ACQUIRE (Adaptive Constraint-based Query Interface) [Huang *et al.*, 2001] is another project making use of the interaction with the user to dynamically build a meta-search engine interface. Interactions with the user can be considered contextual information.

**Interactive Image Retrieval**

Interactive retrieval involves subjective elements such as the user's background, tastes, knowledge and experience and the nature of the task. Although interactive retrieval is a more real-world scenario, it is currently under-researched. The reason for this might be the cost of training and data gathering on a significant and representative user pool. [Goodrum and Spink, 1999] found that interactive search queries involve usually very few search terms, unlike the Adhoc retrieval task of ImageCLEF where a narrative of a few terms is provided (see Section 3.1). Research in interactive retrieval focuses mainly on three areas:

- Query formulation: deals with the final form a query is presented to a retrieval engine
- Result presentation: investigates the presentation of the retrieved results to the user and collecting relevance feedback
- Browsing: the user-interface component allowing the user to perform a new query formulation step based on the results presented to her

## 2.6 Search Behavior

While some studies have focused on analyzing user needs in image archives such as [Armitage and Enser, 1997], on the effectiveness of current image indexing practices [Markkula and Sormunen, 2000] and on work flow-based interface design for image search [Garber and Grunes, 1992], models of search behavior specific to image search are rare in the literature. The purpose of using these models is to relate the different factors involved in information seeking and retrieval. One such model described in [Pharo, 2004] consists of the following categories:

- Work Task: is characterized by the end goal of the search, its complexity and size.
- Searcher: The person carrying out the search as modelled by her knowledge and experience of the work task, of searching in general and of the particular system as well as her eduaction, motivation, tenacity, uncertainty and attention.
- Social/Organizational Environment: The goal of the organization and other people whose opinions/decisions might influence the search.
- Search Task: The information sought after, the strategies employed to achieve the search and the complexity of the task as measured by the subtasks (steps) involved and the predictability of the search.
- Search Process: In this model the search process is divided into situations and transitions, where a situation is the condition of examining a resource to find the sought information while a transition involves searching for these resources.

Of these model elements, the Searcher, the Search Task and the Search Process are of special interest due to their potential tractability. We therefore look into them in more detail.

### 2.6.1 Search Task

A search task differs from the work task in that it is specific to the particular search session in question. According to [Pharo, 2004], the main aspects of a search task is its goal. Search task goals can vary in clarity. Consider the following examples:

- Finding a previously known image to the searcher (as is the case in the interactive track of ImageCLEF).
- Finding an image of a known named entity (for example, a specific person or painting).
- Finding an image of a general entity (for example, an image of a horse carriage in the snow).
- Finding multiple images (for example, showing the subject from different angles or in different situations).
- Finding the best image (which can be a subjective or objective measure).

In the above examples, the goal of the task becomes defined (and possibly redefined) at different stages of the search process. This would likely result in the searcher employing different search strategies and tactics. These are described in the next section.

### 2.6.2 The Search Process

To continue with the model described in [Pharo, 2004], the search process consists of *search situations* and *search transitions*. Moving between search situations (looking through actual information) and transitions (looking through resources) is a search tactic (and if planned ahead, is a search strategy). While clustering results correspond to a search transition rather than a situation, unlike traditional indexes, image cluster results allow more direct relevance feedback.

Search situations and transitions have the following attributes which are potential context indicators [Pharo, 2004]:

- Action: Entering or changing a textual and/or visual query, indicating relevance and exploring clusters.
- Accumulated results: Successful matches until a given moment.
- Accumulated effort: The work put towards finding matches (for example in terms of actions, clicks etc..),
- Time: The total time spent in the search process.

- Relevance Judgment: The degree of relevance of the found matches.

Search tactics include the use of browsing, querying, query reformulation, manual expansion of the query and relevance feedback. [Teevan *et al.*, 2004] distinguishes two search strategies: *orienteering* and *teleporting*. Orienteering proceeds locally through the search process using contextual information without specifying the full need at the beginning while teleporting attempts to reach the target directly through an accurate definition of the information sought. The authors argue that orienteering is more popular and among searchers due to the decreased memory load, a better overall view of the data and understanding of the search results, all of which can also be said of clustering.

An alternative framework for categorizing search behavior based on a search goal hierarchy is found in [Rose and Levinson, 2004]. The highest levels of the hierarchies are three categories:

- **Navigational:** The goal is to go to a specific location.
- **Informational:** The goal is to find information about something.
- **Resource:** The goal is to use the resource in itself.

A relationship is then established between these goals and the user's search behavior as manifested in query formulation, the results and in the user's interaction with the system.

### 2.6.3 The Searcher

The searcher is the user directly interacting with the search system. The searcher's knowledge, which can be broken down into knowledge of the work task, of the search task, of the searching process and of the search system all affect the strategies and tactics adopted by the user. Other personal factors influencing the search process include the searcher's education, her level of motivation to interact with the system and perseverance in achieving subtasks (A study by [Spink *et al.*, 2002] has found that most searchers do not go beyond the first or second page of search results), and the searcher's span of attention.

## Conclusion

This chapter presented the most salient trends and research directions in the image retrieval domain including the main ones employed in this research: text-based-retrieval, content-based retrieval, combination of both modalities and relevance feedback. The next chapter focuses on benchmarking efforts in the domain.

# Chapter 3

# Image Retrieval Evaluation

This chapter introduces the benchmarks, datasets and metrics that are commonly used in the evaluation of image retrieval, including the work presented in this thesis. Section 3.1 introduces benchmarking in image retrieval, Section 3.2 the queries used in the benchmarks, Section 3.3 the collections, Section 3.4 the metrics and Section 3.5 the standard evaluation software.

## 3.1 Benchmarking

An information retrieval benchmark is a framework for the evaluation of IR systems. The essential components of a benchmark are the image collection (Section 3.3), a set of queries or information needs (Section 3.2), and the corresponding relevant images also known as the *ground truth* or *Gold Standard*. To the best of our knowledge **ImageCLEF** is the only existing standard benchmark for measuring the effectiveness of image retrieval systems. ImageCLEF was modeled on the TREC benchmark for Text retrieval [Smeaton, 2001]. The evaluation methodology used in ImageCLEF is described in [Müller *et al.*, 2006]. For its first three years, ImageCLEF used a dataset provided by the St. Andrews Library consisting of 30,000 images for the adhoc retrieval task (see Section 3.3). While the St. Andrews collection has been beneficial in jump-starting an image retrieval benchmark, it has significant drawbacks. Most notably, all images in the collection share the same domain: Scottish historical pictures. Another important disadvantage is the dominance of grey-scale

```
<top>
<num> Number: 1 </num>
<title> accommodation with swimming pool </title>
<narr> </narr>
<image> topics/01/3793.jpg </image>
<image> topics/01/6321.jpg </image>
<image> topics/01/6395.jpg </image>
</top>
```

Figure 5: Example of a topic from ImageCLEFPhoto 2007

images. This is challenging for most content-based systems which rely on color and texture features and does not represent the majority of available digital images.

## 3.2 Queries

This section describes the queries that were used in the ImageCLEFPhoto benchmark and against which this research is evaluated.

### 3.2.1 ImageCLEFPhoto 2007 Queries

In the 2007 ImageCLEFPhoto campaign, sixty queries were provided. The queries are listed in Table 2. The queries were provided in XML format including fields for the topic number and the example images. A narrative field was always left empty. Topic 1 is given as an example in Figure 5. The query implies a search for an image depicting an accommodation (i.e. a lodging such as a home, a hotel, a hostel, or a guest house) with a visible swimming pool in the image. Some of the queries specified if the subject of the search should be in the foreground.

### 3.2.2 ImageCLEFPhoto 2008 Queries

In 2008, the ImageCLEFPhoto campaign decided to use 39 topics out of the 60 used in 2007, in order to adapt the queries to the goal of that year, promoting diversity. Most of the images were annotated with a *title*, a brief semantic and visual *description* and a *notes* field as well as the *location* and *date* of the image, with

some annotations missing in some or all of the fields.

Queries covered a wide compass of semantic and visual difficulty. Visually-oriented queries included *straight road in the USA*, while an example of a semantic query is *views of Sydney's world-famous landmarks*. Answers to the queries were expected in the form of a ranked list of the file names relevant to the query in the TREC-EVAL format.

### 3.2.3 ImageCLEFPhoto 2009 Queries

Unlike the 2007 and 2008 ImageCLEFPhoto campaigns which employed the IAPR TC-12 collection, the 2009 campaign used the Belga news agency collection described in Section 3.3.3. The queries consisted of a *title* and a *narrative*. The narrative is a detailed explanation of results that would be considered relevant to the query, as well as those that would be judged irrelevant.

## 3.3 Collections

One of the early collections used in image retrieval evaluation and specifically CBIR is the Corel photo CDs. Various research projects have used the Corel photo collection including [Qiu, 2004] in addition to [Markkula and Sormunen, 2000]. While this collection is interesting in visual clustering experiments since the ground truth is provided by the original image division, it is not useful for text-based retrieval since the images are not annotated. In addition, the Corel collection is a proprietary one provided on CD for a fee and hence not suitable for research and benchmarking purposes. As discussed in [Müller *et al.*, 2002], the separation of images into substantially different classes makes the Corel dataset too artificial for the purpose of benchmarking image retrieval.

There is growing interest in using collections of images crawled from the web, since these provide a more realistic set. [Li *et al.*, 2005], [Cai *et al.*, 2004b], [Cai *et al.*, 2004a] and [Gao *et al.*, 2005] used different subsets of web images like animal and museum photographs. [Coelho *et al.*, 2004] used images from the Brazilian web (.br domain). Unless frozen in time at a particular instance and saved for future use, crawling the web would produce different results and so would not be appropriate for comparative ends. Other possible

sources of data on the web include public photo repositories such as www.flickr.com, www.webshots.com, www.imagestation.com. It is unclear however how permissions and copyright issues could be handled to use them.

More standard collections are described below.

### 3.3.1 The St. Andrews Collection

The St. Andrews collection of historic Scottish photographs was used for the ImageCLEF Ad-hoc photographic retrieval task between 2003 and 2005. It consists of 28,133 black and white photographs from the library of St. Andrews collection, a collection of mostly black and white Scottish historical pictures [Reid, 1999]. During these years, the focus of the benchmark was Cross Language Image Retrieval (CLIR) with relatively abundant text annotations. The topics were also translated into several languages.

Following is an example illustrating the annotation accompanying an image from the 2003 ImageCLEF campaign [Clough *et al.*, 2005b].

```
Record ID: JV-A.000460
Short title: The Fountain, Alexandria.
Long title: Alexandria. The Fountain.
Location: Dunbartonshire, Scotland
Description: Street junction with large ornate fountain
with columns, surrounded by rails and
lamp posts at corners; houses and shops.
Date: Registered 17 July 1934
Photographer: J Valentine & Co
Categories: [ columns unclassified ][ street lamps - ornate][ electric street lighting ]
[ shepherds& shepherdesses ][ streetscapes ][ shops ]
Notes: JV-A460 jf/m
```

### 3.3.2 The IAPR-TC12 Collection

The ImageCLEF data used in the early years of the benchmark belonged to the same domain (Scottish historical pictures). In recognition of the shortcomings of this closed-domain, and the use of mostly black and white pictures, the organizers of ImageCLEF decided to turn to a more appropriate collection that reflects a wider and more diverse pool of images. The collection of choice was the IAPR-TC12 collection.

The IAPR-TC12 collection was started by the International Association of Pattern Recognition (IAPR). The collection was first described in the ImageCLEF 2005 proceedings consisting of 25,000 annotated thumbnailed images. The images belong to a variety of categories including sports, cities, landscape, animals, people and action shots. A complete description of the collection can be found in [Grubinger *et al.*, 2005].

In the 2006 ImageCLEFPhoto, the collection used in the ImageCLEF benchmark comprised 20,000 color photographs out of the 25,000 described in 2005. The images were annotated with semi-structured captions in German and English. The following is an example of the annotation of an image used in the 2005 benchmark. The corresponding image is shown in Figure 7.

In 2007, the same collection was used again, however, it was not permitted to use the semantic description field. Annotations in Spanish were also provided.

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

```
<THUMBNAIL>thumbnails/16/16019.jpg</THUMBNAIL>
</DOC>
```

### 3.3.3 The Belga Collection

In order to expand the scope of benchmark and improve the level of its diversity, as well as to prevent over-fitting the data, the ImageCLEF 2009 campaign introduced a new dataset, the Belga collection. The collection used for the ImageCLEFphoto 2009 campaign contains almost half a million images annotated with unstructured English annotations describing the image [Paramita *et al.*, 2009]. Structured annotations include fields that break them down by category such as location, date, event, and keywords. Unstructured annotations on the other hand lump the description into one uncategorized field and are more challenging, since they do not lend themselves to database-driven systems with specific fields. The source of the photographs of the Belga collection is Belga [1], a Belgian news agency. The images are of varying dimensions, and are sometimes not orientated correctly. Some of the images are grey-scale while others are color images. An example of the annotation of image 10151 shown in Figure 8 is as follows:

```
<DOC>
<DOCNO>10151</DOCNO>
<DESCRIPTION>A masked man throws a paint bomb towards the Iranian embassy in The
 Hague 11 April, as a policeman runs to arrest him. A group of demonstrators,
 calling themselves the opposition of the Iranian left wing, stated their protest
 action against the violation of human rights in Iran. 21 People were arrested.
</DESCRIPTION>
</DOC>
```

[1] http://www.belga.be

### 3.3.4 The INEX MM Wikipedia Collection

In 2008, ImageCLEF introduced an new image retrieval task, the WikipediaMM Task. The INEX MM Wikipedia image collection of approximately 150,000 images in jpeg and png format was intended for the same retrieval task as the ImageCLEFPhoto task but with a more diverse and noisy collection that resembles retrieval from the web [Tsikrika and Westerveld, 2008]. Unlike the IAPR collection (Section 3.3.2), the Wikipedia image collection includes images of various dimensions including icons and is thus considerably more challenging.

In the experiments presented in this dissertation, only the IAPR-TC12 collection and the Belga collection are used. The reason for not using the St. Andrews collection is that the content-based method relies mainly on color features. In addition, the INEX MM Wikipedia collection was too noisy and required more pre-processing than the method provided by the visual retrieval engine used.

## 3.4 Metrics

In this section, the most common metrics used in the evaluation and benchmarking of image retrieval systems are presented. We divide these into two types of metrics:

- Traditional metrics: those that have been frequently used in Information Retrieval to measure the accuracy (precision) of results, and their prevalence (recall).
- User-oriented metrics: Metrics that are used to capture the diversity of results.

### 3.4.1 Traditional Metrics

This section presents the traditional metrics commonly used in Information Retrieval Benchmarking.

#### Precision

Precision is one of the standard IR retrieval metrics. It reflects the fraction of the retrieved results that are considered relevant to a given query (True Positives) to the total number of results retrieved, as indicated in

Equation 1. Precision can be calculated at a specific cutoff level K, giving rise to measures P@K such as p@10, p@20... These measures capture well real-life scenarios of users' perceived satisfaction, since most users only check the first pages of results presented.

$$Precision = \frac{Number\ of\ Relevant\ Images\ Retrieved}{Total\ Number\ of\ Retrieved\ Images} \tag{1}$$

**Recall**

Recall is another frequently used metric in the IR community. It represents the fraction of the results retrieved to the total number of matching documents in the collection. Equation 2 defines recall.

$$Recall = \frac{Number\ of\ Relevant\ Images\ Retrieved}{Total\ Number\ of\ Relevant\ Images\ in\ the\ Collection} \tag{2}$$

**Precision/Recall**

In a typical information retrieval task *precision* and *recall* figures are inversely proportional. Some metrics, such as the F-measure, attempt to combine precision and recall into a single figure (see Equation 3). A frequently used F-measure is the F1 score which is the harmonic mean of precision and recall as shown in equation 4.

$$F_{\beta} = (1+\beta^2).\frac{Precision + Recall}{(\beta^2.Precision) + Recall} \tag{3}$$

$$F = 2.\frac{Precision.Recall}{Precision + Recall} \tag{4}$$

**Mean Average Precision (MAP)**

The main metric used in image retrieval benchmarking is the Mean Average Precision (MAP), which combines both precision and recall aspects. Average precision for a given query is the average of precision at each of the top recalled documents. The MAP over a set of queries (Q) is then the mean of Average Precision as shown in Equation 5 [Manning *et al.*, 2008].

$$MAP(Q) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} 1/m_j \sum_{k=1}^{m_j} Precision(R_{jk}) \tag{5}$$

**R-Precision**

Another metric that is highly correlated to the MAP is the R-Precision [Aslam *et al.*, 2005]. R-Precision is the precision at rank R, where R is the number of relevant documents.

### 3.4.2 User-oriented Metrics

While the previously discussed measures evaluate retrieval from a system's point of view, measures like *Coverage* and *Novelty* are user-oriented.

**Coverage**

Coverage establishes from a user's point of view the fraction of the retrieved documents already known to the user to be relevant to the total documents retrieved known to the user as per Equation 6.

$$Coverage = \frac{Relevant\_known}{Retrieved\_known} \tag{6}$$

**Novelty**

The other user-driven metric, Novelty, attempts to capture the proportion of relevant documents previously unknown to the user that were retrieved. It is defined in Equation 7.

$$Coverage = \frac{Relevant\_unknown}{Relevant\_unknown + Retrieved\_known} \tag{7}$$

**User Satisfaction**

Similar to Novelty and Coverage, User Satisfaction is a qualitative measure often used in the context of evaluating interactive image retrieval. This can be achieved through questionnaires which poll the users for

their degree of satisfaction with different elements of the system such as speed, ease of use, presentation, and other usability metrics.

**Cluster Recall**

While not strictly a user-oriented metric, Cluster Recall has been used by the ImageCLEF campaigns (2008 and 2009) in order to assess the diversity of results with respect to a known set of clusters for each result. Diversity tasks and metrics in information retrieval aim to promote better coverage of the results, sine queries are often ambiguous, leading to various different semantics. For example, the query topic "Clinton" from ImageCLEFPhoto 2009 was divided into three known clusters: Hillary Clinton, Obama Clinton (a cluster combining Obama and Clinton), and Bill Clinton. Diversity is also used in the TREC Web Track [Clarke *et al.*, 2011a] where it is measured using intent-aware metrics. A comparative analysis of diversity measures used in information retrieval evaluation can be found in [Clarke *et al.*, 2011b]. Cluster recall in ImageCLEF is measured at different levels of recall namely CR@5, CR@10, CR@15, CR@20, CR@30, CR@50, CR@100 and CR@1000, where each measure is calculated by dividing the number of clusters to which the images retrieved at this level of recall belong over the total number of clusters.

$$CR@Recall = \frac{Number\_of\_Clusters\_Discovered\_at\_Recall\_Level}{Total\_Number\_of\_Clusters\_Known} \tag{8}$$

## 3.5 The Evaluation Software: TREC_EVAL

All three ImageCLEFPhoto campaigns in which we participated for the benchmarking of our work employed the TREC-EVAL standard evaluation software[2]. The metrics in TREC-EVAL were selected by NIST (National Institute of Standards and Technology) for evaluation of the TREC (Text Retrieval Conference) campaigns. A discussion of the metrics can be found in [Buckley and Voorhees, 2000].

All trec-eval measures are binary in nature, such that the results are considered either relevant or not relevant. The results are submitted in the trec-eval format which is a ranked list of the results relevant to each

[2]http://trec.nist.gov/trec_eval

query in turn.

## Conclusion

In this chapter, we have described the main datasets used in benchmarking image retrieval, as well as the metrics used which will be used as evaluation criteria of the different methods applied in this dissertation. Evaluation in the nascent image retrieval domain has been over-shadowed by the long tradition of text retrieval evaluation. Precision and Recall figures (Section 3.4) have been criticized even in text retrieval for reflecting related information. Image retrieval evaluation and specifically the *semantic gap* entail a subjective evaluation. [Harper and Hendry, 1997] advocates the us e of *micro-evaluation*, the comparison of different users and searches, rather than *macro-evaluation*, the averaging of results on users. [Huijsmans and Sebe, 2005] proposed employing normalization to take into the consideration the size of both the relevant and irrelevant classes.

While the *Mean Average Precision (MAP)* adopted by the ImageCLEF campaign (Section 3.1) accounts for a certain subjectivity (the relevance of an image is determined by consensus of the group of evaluators), it is impractical in the absence of a significant and diverse pool of judges. The level of detail and explicitness of the annotation of the ImageCLEF data give it an artificial quality and make it prone to over-fitting by the participating systems. An example can be found in [Besançon and Millet, 2005] where the 2004 system did not fit the following year's data. Another drawback of ImageCLEF in interactive retrieval is low participation (on average two groups every year) and the vagueness of the task (in the last interactive ImageCLEF task, the participating groups were required to compare two systems developed by the same group). Despite these disadvantages, ImageCLEF remains the only benchmark available for comparative evaluation in image retrieval.

One problematic area of evaluation is the external-resource inter-dependence. As mentioned in Section 2.1.6, some of the external resources that might be useful in image retrieval include the information retrieval engine, the Query-By-Example module, WordNet and possibly other tools such as a named-entity tagger and a stemmer. Due to these tools' own shortcomings, it is difficult to assess the independent performance of the

system. Moreover, a domino-effect is likely to occur when one tool relies on the invalid output of another.

The next chapter describes and evaluates the single-modal retrieval methods used in this dissertation.

Table 2: Query Topics at ImageCLEF 2007

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

Figure 6: Example of a Query from ImageCLEFPhoto 2008

```
<top>
<num>
 Number: 3
</num>
<title>
religious statue in the foreground
 </title>
<cluster>
 statue
</cluster>
<narr>
Relevant images will show a statue of one (or more) religious figures such as gods,
 angels, prophets etc. from any kind of religion in the foreground. Non-religious statues
 like war memorials or monuments are not relevant. Images with statues that are not the
 focus of the image (like the front view of church with many small statues) are not relevant.
The statues of Easter Island are not relevant as they do not have any religious background.
</narr>
<image> SampleImages/03/31.jpg </image>
<image> SampleImages/03/7446.jpg </image>
<image> SampleImages/03/35577.jpg </image>
</top>
```

Figure 7: Example Image From the IAPR-TC12 Collection

Table 3: Query Topics at ImageCLEF 2009

| ID | Topic | ID | Topic |
|---|---|---|---|
| 1 | leterme | 26 | obama |
| 2 | fortis | 27 | anderlecht |
| 3 | brussels | 28 | mathilde |
| 4 | belgium | 29 | boonen |
| 5 | charleroi | 30 | china |
| 6 | vandeurzen | 31 | hellebaut |
| 7 | gevaert | 32 | nadal |
| 8 | koekelberg | 33 | snow |
| 9 | daerden | 34 | spain |
| 10 | borlee | 35 | strike |
| 11 | olympic | 36 | euro |
| 12 | clinton | 37 | paris |
| 13 | martens | 38 | rochus |
| 14 | princess | 39 | beckham |
| 15 | monaco | 40 | prince |
| 16 | queen | 41 | princess mathilde |
| 17 | tom boonen | 42 | mika |
| 18 | bulgaria | 43 | ellen degeneres |
| 19 | kim clijsters | 44 | henin |
| 20 | standard | 45 | arsenal |
| 21 | princess maxima | 46 | tennis |
| 22 | club brugge | 47 | ronaldo |
| 23 | royals | 48 | king |
| 24 | paola | 49 | madonna |
| 25 | mary | 50 | chelsea |

Figure 8: Example Image From the Belga Collection


Figure 9: Example Image From the INEX MM Wikipedia Collection

# Chapter 4

# Single-Modal Retrieval Methods

While the trend in image retrieval research is to attempt to combine text and image modalities, most of the combination approaches rely on some form of separate textual and visual retrieval components. In such systems, the performance and precision of the individual components often determine the efficacy of the combined approach. This chapter lays the groundwork for the modality fusion strategies that are explored in the next chapter. In particular, the baseline retrieval methods employed in the experiments are described, in order to establish the potential and strength of each modality separately. In addition, different single-modality paradigms are compared to examine the effect of changing these on the results.

Section 4.1 presents and analyzes our text retrieval component, comparing two of the principal Information Retrieval models: the vector-based TF-IDF model and the probabilistic model. An experiment involving pre-clustering of the text annotation and augmenting the results with members of the same cluster is also presented. Section 4.1.7 is an assessment of the effects of the size and nature of the corpora and queries on the results of the retrieval. As indicated in Chapter 1, these are essential characterizing factors of an Image Retrieval system. The size of the corpus in the experiments described here is a function of two criteria: the number of documents contained in the corpus (an image document and any text annotation associated with it is considered as one document), and the size of textual annotation provided. The nature of the corpus refers to its domain and level of generality. Finally Section 4.2 concludes the chapter with a description of the

visual-only retrieval component, including a comparison between a system based on the MPEG-7 standard, and the more basic but efficient block-based method utilized in this work.

## 4.1 A Baseline for Text Retrieval

There are several factors explaining the prevalence of text-based retrieval methods in the Image Retrieval domain on the web. Most notably, the relative speed of text retrieval compared to visual retrieval. Parsing text, preprocessing it, and comparing a text query to the document collection, is substantially faster than analyzing an image to extract visual features from it. Text indices are also much smaller in size than visual ones, and consequently have a much smaller memory fingerprint. Original queries are normally formulated in text, and even when using fusion method such as those described in the next chapter, the text query is often processed first, unless an example image is available. Finally, text-only methods produce by far more precise results compared to visual-only methods.

This section describes the retrieval experiments carried out on the text corpora, without involvement of visual features. For text retrieval, in order to establish a baseline for the vector-based retrieval model, two Java-based information retrieval platforms were used: the Apache Lucene engine [Hatcher and Gospodnetic, 2004], a Java-based text search engine[1], and the Terrier Information Retrieval platform developed at the University of Glasgow [Ounis *et al.*, 2006][2]. Both frameworks implement the TF-IDF paradigm, while the Terrier platform also implements a number of probabilistic similarity measures.

Three measures were compared for the text baseline: TF-IDF, BM25 and PL2. Terrier's TF-IDF implementation uses term frequency defined by Robertson [Robertson and Walker, 1994]. PL2 weighting uses Poisson estimation for randomness, Laplace succession for first normalization, and Normalization 2 for term frequency normalization. The BM25 and PL2 models used are available in Terrier.

[1] http://lucene.apache.org
[2] http://www.terrier.org

### 4.1.1 Text Preprocessing

The first step in text retrieval is the preprocessing stage. Preprocessing refers to preparing the text collection, and the text query, for more efficient indexing and more accurate results. This step is dependent on the collection and especially on its noise level. Noisier collections, with relatively more text in a given document that is irrelevant to the images, require more elaborate preprocessing to remove the irrelevant data that is not related to their visual content. Specific language processing is also often required [Martínez-Fernández *et al.*, 2006], such as removing non-alphabet characters and characters that belong to a different language. In annotated image collections, preprocessing often includes removing the annotation tags before indexing the collection, as well as query tags from the query collection.

The most common preprocessing steps are stop words filtering and stemming or lemmatization. Removing stop words, grammatical words that do not contribute to the retrieval process given their very high rate of occurrence, can dramatically decrease the size of the index, as well as improve the accuracy of results. In our experiments, stop words include all the words in closed word classes (determiners, prepositions, conjunctions and pronouns). They also include some open-class items (nouns, verbs, adjectives, adverbs), which normally do not contribute to the semantics. For example, primary helping verbs (e.g. be, have, do) and modal helping verbs (e.g. can, will, shall) are excluded. Other stop words include common adjectives, and some temporal, locative and quantitative adverbs (e.g. here, mostly). In the indexing phase, the stop words list used was the internal one used in Terrier, while in retrieval some additional words specific to the collection were added to prevent the introduction of noise in query expansion as discussed in section 4.1.3. For the full list of stop words used, see Appendix A.

Another common text preprocessing step is stemming or lemmatization, which increases the probability of matching a query where the word has a different inflection from the one in the corpus. An inflection is a different form of the word that does not significantly modify its semantics. Stemming is an approximation of the morphology of the word. By contrast, lemmatization incorporates syntactic knowledge, such as parts of speech, in order to determine a more accurate morphology. In [Juffinger *et al.*, 2009], using lemmas instead of the exact word form was found to increase the Mean Average Precision (MAP) (discussed in Section 3.4.1)

by 1.3%. The better accuracy of lemmatization is at the cost of speed and computing resources. In the experiments presented here, only stemming was used. The stemming performed in the experiments employed the internal Terrier Porter stemmer in the indexing of the collection, and the Snowball stemmer [Porter, 2001], also based on the Porter stemmer, for stemming the query.

### 4.1.2 Indexing the Document Collections

Before the text can be indexed, it is necessary to prepare it using the preprocessing steps as described in Section 4.1.1. When indexing using *Terrier*, the option of block-indexing for phrase querying was applied. Query terms are considered unioned by Terrier in order to promote recall. The rest of the terms are converted to lower case and used in indexing. The *title*, *notes*, *location* and *description* fields (see Section 3.3) of the documents were indexed. Metadata including the document header, the document number, the example image file name, and that of its thumbnail were excluded. The Terrier collection class used is the SimpleXMLCollection, which handles TREC-like collections on condition of having valid XML. A problem with the SimpleXMLCollection class was rectified in the course of this dissertation, and contributed back to the open source project[3]. The result of the indexing step is the production of an inverted index, which is accessible through an API. In the experiments with the Lucene engine, stop-words were also removed, and the data was indexed as *field data* retaining only the *title*, *notes* and *location* fields, all of which were concatenated into one field.

### 4.1.3 Processing the Queries

As in the case of the document collection, queries are tokenized and preprocessed similarly; stop words and punctuation are removed and the rest of the terms are stemmed. For stripping the query of stop words, a custom stop word list was used to adapt to the nature of the collection which consists of news snippets. It includes, in addition to the aforementioned categories, a few more collection-specific terms, for example, “Belga”, the name of the press agency which figures in almost all documents as the source of the news. Similarly, the names of the days of the week and the months are also excluded since they figure in the header

[3] http://terrier.org/forum//read.php?3,1065,1066#msg-1066

```
<num> Number: 6 </num><title> straight road in the USA </title>
<cluster> state </cluster><narr> Relevant images will show a straight road or highway
(either empty or with traffic) in the United States of America. A road is considered
 to be a straight road if there is no curve visible in the image. Images with roads
with a curve are not relevant. Images with straight roads that are not in the USA are
 not relevant.  Images with roads too short to determine whether they are straight
or not (like side views) are not relevant. </narr>
<image> SampleImages/06/37537.jpg </image>
<image> SampleImages/06/37736.jpg </image>
<image> SampleImages/06/37754.jpg </image>
</top>
```

Figure 10: Original Query for Topic 6 of ImageCLEFPhoto 2008

of the news snippet. A full list of the stop words used in the experiments can be found in Appendix A. The rest of the terms are converted to lower case and stemmed using the Snowball stemmer [Porter, 2001]. The Snowball stemmer is based on the Porter algorithm and available under the BSD license[4].

The queries consist of a *title* and a *narrative*. The narrative is a detailed explanation of results that would be considered relevant to the query, as well as those that would be judged irrelevant. When constructing the query, named-entities are given more weight, and multiple-token named-entities are chunked into one term by adding quotes around them. Named-entities are recognized using simple capitalization heuristics. Negative sentences of the narrative, which indicate irrelevant criteria, are identified using a negative keyword list. They are then discarded so as to avoid the cost of extensive logical and semantic processing. All text query terms were explicitly joined using the *OR* operator in the experiments using the Lucene engine, while the disjunction is implicit by default in Terrier. Figure 10 illustrates the unprocessed topic number 6 used in ImageCLEF 2008.

As we have seen in Section 4.1.7, using the first sentence of the narrative field (narr) which expresses positive examples, in addition to the title field, improves the result. By contrast, the rest of the narrative, like negative examples, needs semantic processing to avoid introducing noise. Hence, only the sentences conveying positive sentiment of the narrative field are retained in the query. The sentences containing negation words, in the provided example “no” and “not”, are excluded. Stop words are removed, and the rest of the tokens are stemmed. The example image fields (*image*) are metadata which is also excluded. In addition, a

[4]http://snowball.tartarus.org

Table 4: Comparison between TF-IDF and Probabilistic Models on the IAPR TC-12 Text-Only Retrieval

| Model | MAP | P10 | P20 | P30 | Relevant |
|---|---|---|---|---|---|
| TF-IDF | 0.3390 | 0.5000 | 0.4385 | 0.3752 | 1883 |
| BM25 | 0.3393 | 0.5026 | 0.4372 | 0.3786 | 1868 |
| PL2 | 0.3335 | 0.4923 | 0.4333 | 0.3752 | 1868 |

Table 5: Comparison between TF-IDF and Probabilistic Models on the Belga Text-Only Retrieval

| Model | MAP | P10 | P20 | P30 | Relevant |
|---|---|---|---|---|---|
| TF-IDF | 0.5124 | 0.6800 | 0.7870 | 0.7847 | 19969 |
| BM25 | 0.5099 | 0.7680 | 0.7870 | 0.7867 | 19743 |
| PL2 | 0.5146 | 0.7780 | 0.7790 | 0.7840 | 20095 |

weighting parameter is added to USA, the original named entity in the title of the query. An important step of this phase is assigning weight to the query tokens. Since the query consists of both a title and a narrative, it is presumed that the title contains more terms with higher confidence than the narrative. For this reason, the stop words-filtered title tokens are given a higher weight than the other terms in the narrative, unless these appear in the title as well. For example for the query of Figure 10, the final text query sent to Terrier for processing is as follows:

```
usa straight road highway traffic Unite State America
```

### 4.1.4 Probabilistic vs. Vector-based Models Experiments

In the scope of this dissertation, the two main paradigms in text retrieval that were experimented with were the vector-based model and the probabilistic model. The difference between the vector-based model and the probabilistic model on both data sets in single-medium as well as mixed-media experiments was not found to be significant. This is most likely attributed to the scarcity of the text content. The vector-space based model is the TF-IDF model while the probabilistic models experimented with are the Okapi BM25 and PL2 weighting functions. Tables 4 and 5 illustrate the comparison of the results obtained using each model on the IAPR collection and the Belga collection respectively. Section 5.4 will present the same comparison when applied with fusion methods.

### 4.1.5 Enhancing the Text Baseline

Three methods were attempted to enhance the text-only baseline.

1. Adding extra weight to the more relevant terms of the query.
2. Excluding terms that are presumed non relevant according to the narrative.
3. Expanding the query by adding new relevant terms which do not appear in the original query title or narrative.

**Adding extra weight to the more relevant terms of the query:** The idea behind adding more weight (enhancement 1 above) is to favor terms directly related to the query over those that should carry less bearing on the results, especially terms from the narrative. For this reason two criteria were considered:

- Whether the term is a named-entity
- The position of the term in the query

Adding weights is possible in Terrier due to its rich query language. To add more weight to a query, the caret character is used after the term which weight is desired to be altered, along with a decimal weight. For example, to assign twice the weight to the term *USA* in the example used in Section 4.1.3, it is changed to USA^2 and the whole query becomes:

```
usa^2 straight road highway traffic Unite State America straight road highway traffic
Unite State America.
```

The weights experimented with ranged from the neutral weight 1 up to 12 times that weight. It was observed that assigning the same weight to all query terms resulted in a significant degradation of both the precision and recall of the results as illustrated in Table 6. Also, assigning more weight to the first term resulted in lowering the precision and recall as shown in table 7.

Table 6: Assigning All Terms the Same Weight

| Model | MAP | P10 | P20 | P30 | Relevant |
|---|---|---|---|---|---|
| Probabilistic (PL2) | 0.2419 | 0.3410 | 0.3179 | 0.2786 | 1293 |

Table 7: Assigning More Weight to First Term

| Weight | MAP | P10 | P20 | P30 | Recall |
|---|---|---|---|---|---|
| 2 | 0.3278 | 0.4872 | 0.4141 | 0.3632 | 1833 |
| 3 | 0.2946 | 0.4487 | 0.3846 | 0.3427 | 1788 |
| 7 | 0.2318 | 0.3026 | 0.2718 | 0.2641 | 1732 |

**Excluding terms that are presumed non relevant according to the narrative:** Similarly, excluding the terms that appeared in the negative examples given in the narrative (enhancement 2 above) resulted in slightly worse results. The negative sentences provided in the narrative call for more fine-grained language understanding techniques for proper interpretation. Adding synonyms to the query reduced the precision of the results.

**Expanding the query by adding new relevant terms which do not appear in the original query title or narrative:** This is the method that improved the results over single-modal retrieval methods and is explored more in Chapter 5.

### 4.1.6 Textual Clustering of the Collection

Another method for text-only retrieval, pre-clustering of the text collection, was investigated in the context of our participation in the 2007 ImageCLEF photographic ad-hoc retrieval task. The task deals with answering 60 queries of variable complexity from a repository of 20,000 photographic images in the IAPR TC-12 collection. A full description of the task and the collection can be found in Section 3.3.2. Given the small number of relevant results per query, clustering the collection is a possible method for augmenting the retrieved results. Six runs were submitted (see Table 8), aiming to evaluate the text and content-based retrieval tools in the context of the given task. The other purpose of our participation was to experiment with applying clustering techniques to this task, which has not been done frequently in previous editions of the ImageCLEF

Ad hoc retrieval task. While not intended for the evaluation of interactive methods, this task of ImageCLEF could still be useful in the evaluation of certain aspects of such methods such as the validity of the initial clusters. Three of the submitted runs utilized pre-clustering of the data collection to augment the result set of the retrieval engines.

**Clustering in Image Retrieval**

Clustering, as an unsupervised machine learning mechanism, has not been often investigated and benchmarked within the context of ad-hoc image retrieval. This could be due to that clustering methods lend themselves more readily to interactive tasks and iterative retrieval. The ImageCLEFPhoto retrieval benchmark introduced the notion of different semantic clusters for the years 2009 and 2010. In the Information Retrieval field, clustering has been experimented with extensively [Manning *et al.*, 2008]. Its different applications involve clustering the whole data collection, part of it or clustering only the search results. In [Sunayama *et al.*, 2004], images are clustered using labels from the surrounding HTML text.

**Clustering Experiment**

For retrieval, two publicly available libraries were used; *Apache Lucene* [Hatcher and Gospodnetic, 2004] for text and *LIRE* [Lux and Granitzer, 2005] for visual retrieval. Some of the annotation was provided in multiple languages, however since the runs involved only English/English and Visual queries, no translation was employed.

A simple one-pass clustering algorithm was employed, which relied on forming clusters of the terms in the documents as they were processed. If a document's similarity to a cluster exceeded a certain threshold *(n)*, this document and its new terms were added to the term/document cluster. When a document was not associated with any cluster, it was temporarily assigned its own, which was deleted in the end if no other documents were associated with it. Also clusters larger than size *(s)* or smaller than size *(m)* were discarded since they were deemed inconsequential. We did not, however, experiment with the parameters $s$ and $n$ and chose them with the little intuition we had about the data. The resulting clusters overlapped and did not cover all documents.

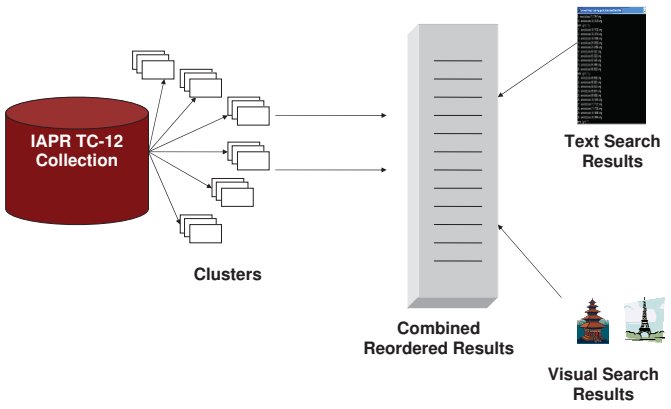


Figure 11: Overview of the Pre-clustering System Used in ImageCLEF 2007.

The following parameters were used in the experiments:

- Number of top results used for cluster expansion $t = 20$
- Number of results retained from image search $r = 20$
- Minimum number of common words to be in the same cluster $n = 3$
- Minimum size of cluster $m = 3$
- Maximum size of cluster $s = 300$

Figure 11 shows an overview of the system used in the 2007 ImageCLEF Ad-hoc retrieval task. In the mixed run (clacTXCB), we combined the results from the Lucene text search and the LIRE visual search by ranking the common ones highest, followed by all other text results and finally the visual results are added at the bottom of the list. This is due to the higher confidence we had in the text search results.

For augmenting the results from the clusters, we searched the clusters for the top $t$ results and whenever one was found we inserted the other members of the cluster at this position in the result set, taking care not

Table 8: Experiments with Clusters at ImageCLEF 2007

| Experiment | Modality | MAP | P10 | P20 | P30 | GMAP | Rel |
|---|---|---|---|---|---|---|---|
| Text+Visual | Mixed | 0.1667 | 0.2750 | 0.2333 | 0.1599 | 0.0461 | 1763 |
| Text+Visual+Clusters | Mixed | 0.1520 | 0.2550 | 0.2158 | 0.1445 | 0.0397 | 1763 |
| Text-only | Text | 0.1355 | 0.2017 | 0.1642 | 0.1231 | 0.0109 | 1555 |
| Text+Clusters | Text | 0.1334 | 0.1900 | 0.1575 | 0.1205 | 0.0102 | 1556 |
| Average run | N/A | 0.1292 | 0.2262 | 0.1913 | 0.1724 | 0.0354 | 1454 |
| Median run | N/A | 0.1327 | 0.2017 | 0.1783 | 0.1659 | 0.0302 | 1523 |
| Best run | Mixed | 0.3175 | 0.5900 | 0.4592 | 0.3839 | 0.1615 | 2251 |

to include duplicate results from different clusters.

**Pre-Clustering Results**

The following six experiments were attempted at ImageCLEF 2007:

1. Text-only: uses Lucene for text search
2. Visual-only: uses LIRE for visual search
3. Text+Visual: combines the results from Lucene and LIRE
4. Text+Clusters: augments clacTX with clusters
5. Visual+Clusters: augments clacCB with clusters
6. Text+Visual+Clusters: augments clacTXCB with clusters

Table 8 shows the results that the six runs obtained at ImageCLEF, as well as the average, median and best runs of the track for reference. Our highest ranked run, (text+visual), is the one that combined results from Lucene (text retrieval) and LIRE (visual retrieval), getting a higher MAP (0.1667) which is significantly better than the text-only run (MAP=0.1355). The significance of this result is confirmed using the Wilcoxon signed-rank test ($z=4.42$, $p<0.0001$)[5]. In addition, the combined run has better performance on all other measures than the other runs. For this run, we used a combined list of the results from both engines, ranking common results highest on the list as described in Section 4.1.6. The poor performance of our text-only

[5]tests conducted using the online tool at http://vassarstats.net/wilcoxon.html

run (MAP=0.1355) can be mainly attributed to the absence of stemming and query expansion/feedback. Indeed, the total number of terms in the text index is 7577. When using a stemmer, this figure is reduced by approximately 800. The results improve by an order of 1% to 2%. As for query expansion, we estimate that the results can improve significantly by employing geographical gazetteers as well as synonyms. Indeed, further examination of the results shows that our poorest results were obtained for queries that reflect a combination of these two factors. For example, the poorest precision of the text-only run was obtained for topics no. 40 *(tourist destinations in bad weather)* and 41 *(winter landscape in South America)*.

The simple method of augmenting results using the pre-clustered data deteriorated the results in all three cases: text, visual and their combination. The main reason is that our clusters were less fine-grained than the requirements of the queries. We retained only 84 clusters of which only a handful were useful. When we experimented with the parameters we found that basing the clustering on a higher number of common words would lead to improving the results over the runs that do not employ the clusters. The one-pass clustering algorithm was unable to find this optimal parameter.

As for the other parameters described earlier, they did not count for significant changes in the results. The number of results retained per visual query (=20) was found to be the most appropriate. Increasing or decreasing it degrades the precision. The same observation applies to the number of top documents (=20) used in augmenting the results, which can be attributed to the degrading precision after the top 20 as can be seen from the results. For the size of clusters we noted that very small clusters, which number below 30, were not useful since it is rare that one of their members happens to be in the query results. On the other hand, large clusters (with size $> 200$) introduce noise and reduce the precision.

### 4.1.7 Effects of the Size and Nature of the Collection

The size and nature of an image collection are two variables that can potentially affect the performance of an image retrieval system. In order to understand the effects of these factors, this section pinpoints significant differences between the datasets and queries used in the experiments, describing the implications of these

differences on the performance of the retrieval methods used. A data collection in the context of this dissertation refers to both the visual and textual components of the data set. The visual component is the image itself, while the textual component is any text attached to the image such as metadata and annotation. Two data sets were experimented with in this research: the IAPR TC-12 and the Belga collection (both were described in Chapter 3).

Table 9 illustrates the major differences in the textual and visual content of these corpora. The characteristics of a corpus that affect the retrieval mechanism include the number of documents in the corpus, the overall number of tokens in the collection, the average length of a document, and the number of unique terms used in it.

Table 9: Datasets

| **Dataset** | **Documents** | **Unique Terms** | **Tokens** | **Average Document Length** |
|---|---|---|---|---|
| IAPR TC-12 | 20,000 | 6,660 | 187,507 | 9.37 |
| IAPR TC-12 (with description field) | 20,000 | 8,036 | 447,603 | 22.38 |
| Belga | 499,998 | 212,395 | 1,844,0128 | 36.88 |

The first criterion, the number of documents in the collection, has important implications for the method used for data modeling and representation, particularly of its visual aspects. The larger an image collection, the less discriminating the features used in visual similarity become. This is illustrated in Section 4.2 by the results obtained using visual similarity on the two data sets. Elaborate visual features require intensive computational resources, often rendering their extraction and the computation of similarity to example images in real time unfeasible for most practical applications. Another challenge that detailed features pose is the size of the index. Huge indices must often be distributed over more hardware to accommodate their size.

Section 4.1 presented the details of the text-only retrieval methods we used. For the textual retrieval component, the second characteristic, the collection's total number of tokens, throws a light on the efficacy of text-only searches on the textual component of the data. Semantically rich collections with significant text content lend themselves naturally to such searches. This explains the better results achieved by the text-only retrieval method on Belga, the data set with more textual content, compared to the IAPR collection (see Sections 5.2.1 and 5.2.2).

Another essential aspect of the corpus in determining the viability of limiting the search process to text-only modality, is the average length of a document in word tokens. In general, smaller documents do not provide sufficient information for precise results when employing text-only retrieval techniques due to data sparseness. While both data sets included only some descriptive metadata, the Belga set had an average of 36.88 terms compared to 22.38 for the IAPR set with the narrative field included.

The final parameter of an image collection could shed a light on its characterization on the generality/specificity scale. The number of unique terms can help determine the extent of the domain variability of the collection. Datasets which are richer in terms of vocabulary entail more domain variability, and call for more generalized retrieval methods. Offsetting this factor is the level of polysemy of the vocabulary.

**The effect of available text on the results**

In order to evaluate the effect of the available text on the precision and recall of the retrieved results, several runs were submitted to the ImageCLEFPhoto 2008 benchmark [El Demerdash *et al.*, 2008]. These runs experimented with changing either the amount of text used from the query, or from the collection. For the query, the *narrative* field was in turn included and excluded from the query. While the narrative field provides more textual information, it is more prosaic in nature and thus not deemed as relevant as the *title* field. For the collection, the *description* field, was excluded from all runs except one.

Following is the list of the experiments performed and their description.

1. Text(title): Uses text search only on the *title* field (Lucene)
2. Visual: Uses visual-only search (block-based method)
3. Text(title+ narr.1): Combines text search on the *title* field and the first sentence of *narrative* field with the text from the first result of the visual search (Pseudo-relevance feedback)
4. Mixed(title+narr.): Combines Visual and Text(title+narrative)
5. Text(title+narr.): Uses text search on *title* and *narrative* fields
6. Mixed(title+narr.+Vis.): Title and narrative combined with visual results

Table 10: Experiments with the Effects of the Size of the Collection (ImageCLEFPhoto 2008)

| Experiment | Modality | MAP | P10 | P20 | P30 | GMAP | Rel | F-measure |
|---|---|---|---|---|---|---|---|---|
| Text(title) | Text | 0.1201 | 0.1872 | 0.1487 | 0.1462 | 0.019 | 1155 | 0.1741 |
| Text(title+narr.1) | Text | 0.2577 | 0.4103 | 0.3449 | 0.3085 | 0.1081 | 1859 | 0.3290 |
| Mixed(title+narr.) | Mixed | 0.2622 | 0.4359 | 0.3744 | 0.3308 | 0.1551 | 1630 | 0.3546 |
| Text(title+narr.) | Mixed | 0.2034 | 0.3205 | 0.2705 | 0.2487 | 0.0780 | 1701 | 0.2875 |
| Mixed(title+narr.+vis.) | Mixed | 0.218 | 0.4026 | 0.3269 | 0.2855 | 0.1290 | 1546 | 0.3384 |
| clacDesc(title+narr.1) | Mixed | 0.3419 | 0.5051 | 0.4256 | 0.3726 | 0.1794 | 2401 | |

7. clacDesc(title+narr.1): includes the *description* field

The results obtained, as well as the track's average, median and best results are shown in Table 10. Manual runs involve human intervention, while automatic ones do not. As Table 10 shows, the runs that used more of the available textual data, clacNoQE, clacTxNr and clacDesc and obtained significantly higher MAP, recall and F-measure than the one that used the title only (clacTX). In addition, it can be observed that omitting the *description* field had a significant negative impact on the precision.

## 4.2 A Baseline for Content-based Retrieval

Despite the challenges facing content-based methods for Image Retrieval, especially in terms of resources and overcoming the semantic gap, there have been many successful attempts to incorporate them in the retrieval process for improving the results. This section introduces the content-based methods employed in the experiments, before fusion with the text. First, experiments using MPEG-7 descriptors are presented, followed by the block-based techniques underlying the visual retrieval engine implemented as an alternative to the MPEG-7 descriptors, and the results obtained from running them over the same benchmarks used in the text-only retrieval methods presented in Section 4.1.

### 4.2.1 MPEG-7 Descriptors

The initial visual retrieval experiments conducted for the research presented, took place in the context of the 2007 ImageCLEF Ad-hoc retrieval task. We experimented with the MPEG-7 descriptors for visual retrieval.

Table 11: Baseline Experiment with Visual Retrieval (ImageCLEF 2007.

| Experiment | Modality | MAP | P10 | P20 | P30 | GMAP | Rel |
|---|---|---|---|---|---|---|---|
| Pure visual retrieval (LIRE MPEG-7) | Visual | 0.0298 | 0.1000 | 0.1000 | 0.0584 | 0.0058 | 368 |
| Visual retrieval + clusters | Mixed | 0.0232 | 0.0817 | 0.0758 | 0.0445 | 0.0038 | 386 |

In order to do that, Version 0.4 of the LIRE library, a part of the Emir/Caliph project available under the GNU GPL license, was employed. At the time of carrying out the experiments, LIRE offered three indexing options from the MPEG-7 descriptors: ScalableColor, ColorLayout and EdgeHistogram (a fourth one, Auto Color Correlogram, has since been implemented). The first two of these are color descriptors while the last is a texture one. All three indices were used in the experiments. The details of these descriptors can be found in [Martínez, 2004]. Only the best 20 images of each visual query were used. The visual queries in that year consisted of the three images provided as example results. Thus, a maximum of 60 image results from visual queries for each topic were used in the evaluation.

Table 11 shows the results the visual runs obtained at ImageCLEFphoto 2007 using the MPEG-7 descriptors. One of the runs uses pure visual retrieval, while the other run augments the results of the first with results from the text clusters that the images belong to. These results clearly show the inadequacy of MPEG-7 features for this task. They represent a baseline to compare to our approach. The next section presents a low-cost alternative to these MPEG-7 descriptors, which significantly outperforms them, in all metrics, but most significantly the precision at the highest recalled documents, which enables the use of the fusion methods described in Chapter 5.

### 4.2.2 Visual Retrieval Using Block-based Techniques

A content-based retrieval engine can also be described as a Query-By-Example (QBE) module. Given an example of a relevant image to a query, the task of the QBE module is to find visually similar images from a collection. While the ultimate aim of content-based retrieval methods is to fetch the most relevant results without human intervention, the current status of hardware limitations on possible image analysis confines this goal to enhancing the results obtained from the textual retrieval component. This is often achieved

through the use of query expansion or some other media fusion method. The goal of our content-based module in concrete terms is to maximize the precision of the results, particularly of the highest ranked documents returned from querying-by-example. While the example images in the experiments described here were provided by the ImageCLEF benchmark, they could also be obtained using relevance feedback mechanisms in actual search tasks.

Since the experiments with MPEG-7 features described in Section 4.2.1 proved them inadequate for use in fusion techniques with textual data, block-based techniques, which have been extensively used in image retrieval, were selected as an alternative for this task. Examples of using block-based techniques in visual retrieval can be found in [Han and Huang, 2005] and [Takala *et al.*, 2005]. These fast and simple techniques are based on partitioning the image into blocks, then performing feature extraction on each block independently. The reason for favoring this approach is its suitability to general, non domain-specific photographic databases, such as the ones used in the ImageCLEFPhoto track [Arni *et al.*, 2008 printed in 2009], where there is not enough information to correctly and meaningfully segment the images. The content-based module was implemented using the Java Advanced Imaging (JAI) API [6].

### 4.2.3 Preprocessing

This section illustrates with images the preprocessing steps that the image collection as well as the example images undergo before feature extraction. Figure 12 shows an example of an image from the IAPR TC-12 collection before pre-processing and indexing. The RGB color histogram of the original image is shown in Figure 13. The image is first converted to the Intensity/Hue/Saturation (IHS) color space, a perceptual color space which is more intuitive and reflective of human color perception than the RGB color space. This also allows for assigning more weight to the hue component which is a better discriminating feature as shown in [Stricker and Orengo, 1995]. The optimal weight of the *Hue* feature was empirically found to be three times the weight of the other features. An RGB representation of the image after the transformation to the IHS color space is shown in Figure 14. The histogram of the IHS image is shown in Figure 15.

The next step in preprocessing the image is applying a median filter to it. The median filter helps in

[6] http://java.sun.com/javase/technologies/desktop/media/jai/

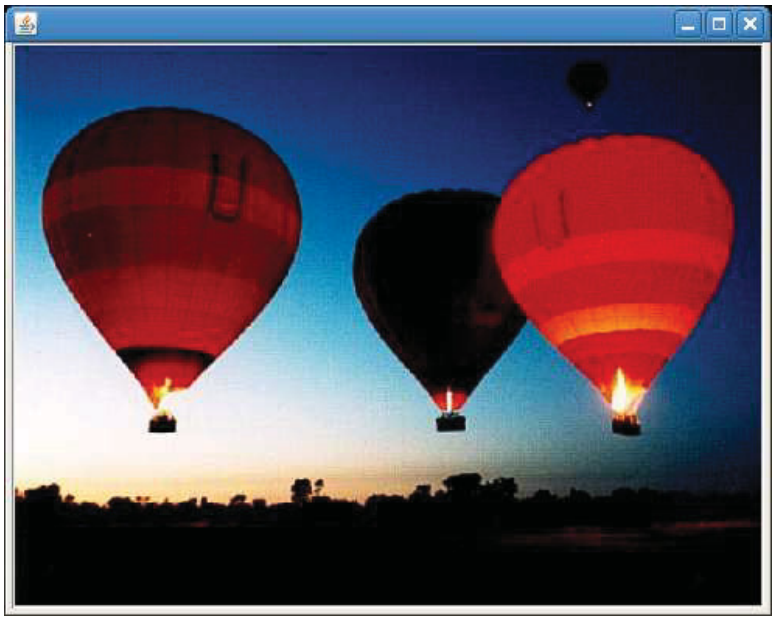

Figure 12: Original Image in RGB Model

removing the noise from the image by eliminating pixels and lines that are outliers. A mask is used to scan the image and the center pixel of the mask is replaced by the median of the mask. A square shape was selected for the mask as shown in Figure 16[7]. The median image is shown in Figure 17 for the RGB color model and in Figure 18 for the IHS model. The histogram of the median image in the IHS model is shown in Figure 19.

### 4.2.4 Extracted Features

Figure 20 shows the different regional divisions used to analyze an image. In order to capture different levels of basic global and local color, texture, and shape information using a block-based method, the image is divided into 2X2, 3X3, 4X4 and 5X5 blocks, yielding 4, 9, 16 and 25 equal partitions respectively. Using finer granularity for partitioning is possible, although at the cost of execution time and storage space. Experiments on the IAPR TC-12 collection yielded a slight deterioration of the results when the next level of division (6X6) was added. This can be attributed to the very small size of the partition. Images in the collection are 480X360 pixels. These divisions, as well as the image as a whole and a center block occupying half the image dimensions, constitute the regions of interest of the image, from which the features are extracted.

[7]Filter API: http://download.java.net/media/jai/javadoc/1.1.3/jai-apidocs/javax/media/jai/operator/MedianFilterDescriptor.html

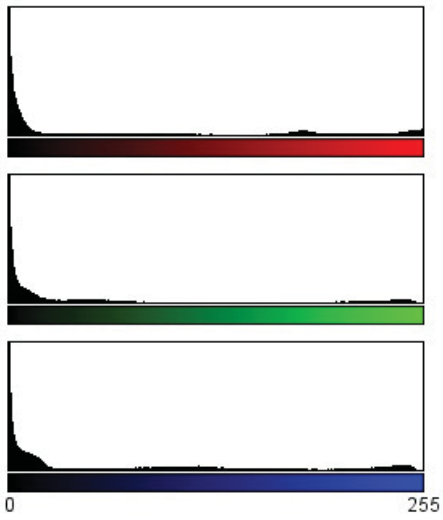


Figure 13: Color Histogram of the Image of Figure 12 in RGB Model

The following features are extracted from each region of interest:

1. A three-band IHS color histogram for each division

2. A histogram of the grey-level image

3. A histogram of the gradient magnitude image for each of the divisions of the grey-level image

4. A three-band color histogram of the image thumbnail

The first feature captures the distribution of the color characteristics of the image, while the grey-level histogram conveys texture information. The grey-level image is shown in Figure 21, and its histogram in Figure 22. The gradient magnitude adds the outline of the shapes in the image. Figure 23 illustrates the gradient magnitude transformation of the image and Figure 24 its histogram. Finally, the thumbnail represents a visual summary of the image. Figure 25 shows the divisions used combined with the feature images.

As has been illustrated before in [Mandal *et al.*, 1996], the moments of histograms are efficient approximations of the entire histogram. Therefore, for each band of each of the histograms, the first two moments (the mean and the average energy) as well as the standard deviation are stored in the index. Moreover, the

Figure 14: Image of Figure 12 in IHS Model

minimum error threshold is used. This feature calculates the threshold which minimizes misclassification error of the histogram modeled as the sum of two Gaussian distributions.

The extracted features yield a feature vector consisting of three statistics for each Region Of Interest (ROI) of each image band. In total, 686 features were used. Table 12 shows the distribution of the 686 feature vector. There are three statistics per band per region of interest. For example, as Table 12 shows, 36 features are used for the 2X2 IHS division (3 bands), and 27 features for the 3X3 gradient Magnitude (1 band) histograms. These are calculated as follows:

2X2 IHS division = 4 Regions Of Interest X 3 bands (IHS) X 3 statistics = 36 features.

3X3 gradient Magnitude division = 9 Regions Of Interest X 3 statistics = 27 features.

### 4.2.5 Visual Retrieval

In the retrieval step, the different partitions of each image in the collection are compared to their counterparts in the query images. Although this simple method does not account for translations and rotations in the image, it is a reasonable choice, especially in the case of photographic images and outdoor images which account for a significant proportion of photographic collections. With the exception of the weight assigned

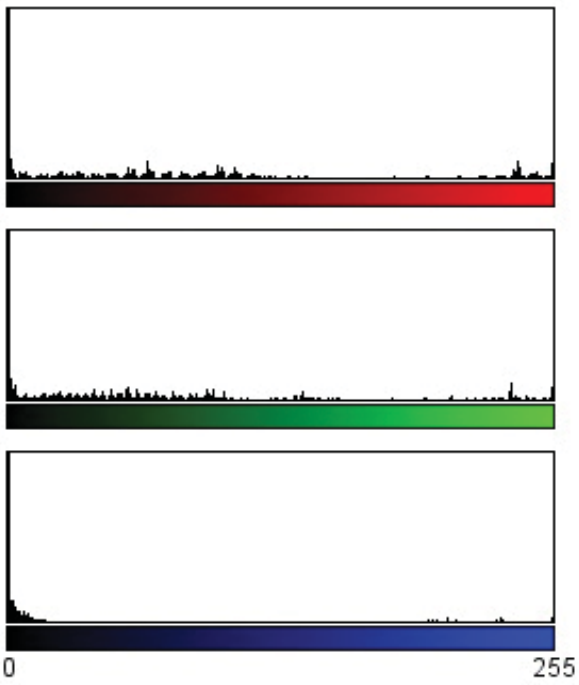


Figure 15: IHS Histogram of the Image Shown in Figure 12

to the hue component of the color histogram (three times the weight of other features), the same weight was assigned to all other features to avoid over-fitting the data.

For the distance measurement, after investigating several measures including the Euclidean and the Mahalanobis distances (see Section 4.2.6) using the IAPR TC-12 collection, the Manhattan distance ($L^1$ Norm) was selected combined with a measure of the number of blocks within a distance threshold. Since all features were represented as histograms with the same number of bins (256), no normalization was necessary. The images in the database were ranked according to their highest proximity to any of the three query images. This choice presumes that the simple features used do not perform equally well on all example images.


Figure 16: Median Mask: All Pixels in the Square (represented by X's) are Used in Calculating the Median

### 4.2.6 Distance Measures

Two types of distance measures have been employed in the visual retrieval component. The first is a traditional metric distance measure, while the second is a quantification of the similarity between images. For the first category of distances, several metrics were compared, including Manhattan distance, normalized Manhattan distance, Euclidean distance, Mahalanobis distance and Bray-Curtis distance. The formulae for the distance metrics experimented with are shown in Equations 9, 10, 11, 12 and 13, where x and y are two corresponding data points in the visual descriptors, and S (in Equation 13) is the covariance matrix. The second type of distance measure used consists in measuring the number of features that fall within an empirically determined threshold.

**Manhattan Distance**

$$d(x,y) = \sum_{i=1}^{m} |x_i - y_i| \tag{9}$$

**Normalized Manhattan Distance**

$$d(x,y) = \sum_{i=1}^{m} \frac{2|x_i - y_i|}{x_i + y_i} \tag{10}$$

Figure 17: Median Image in RGB Model of the Image of Figure 12

**Euclidean Distance**

$$d(x,y) = \sqrt{\sum_{i=1}^{m}(x_i - y_i)^2} \tag{11}$$

**Bray-Curtis Dissimilarity**

$$d(x,y) = \frac{\sum_{i=1}^{m}|x_i - y_i|}{\sum_{i=1}^{m} x_i + y_i} \tag{12}$$

**Mahalanobis Distance**

$$d(x,y) = \sqrt{(x-y)S^{-1}(x-y)} \tag{13}$$

The Manhattan distance was found to yield the best precision among the other metric measures tested on the IAPR TC-12 collection.

### 4.2.7 Impact of the Visual Features on Retrieval

The rank of a matrix calculated using the Singular Value Decomposition (SVD) method is equal to the number of non-zero singular values. Calculating the rank of the matrix of visual descriptors on the IAPR TC-12 visual

Figure 18: Median Image of Figure ]refrgb in IHS Model

features confirmed that most of the features are indeed essential. The rank of the matrix was found to be 485 for the IAPR-TC 12 collection. However when applying SVD on the Belga set, the rank was found to be only 16. We deduce from this that only a few features remained relevant and discriminatory in a much larger dataset. Table 13 shows the effect of removing color distance, all color features, and the gradient feature from the feature vector respectively. We deduce from this table that the color features account for the most important part of the similarity feature vector.

In this chapter, we have described the single-modal retrieval methods,text-only (Section 4.1) and visual-only (Section4.2). We have studied their potential in the retrieval process independently of each other. We have seen how poorly visual retrieval methods perform (Table 11), a minimal difference between vector-based and probabilistic models in text retrieval (Tables 4 and 5), and how the amount of text available influences the textual retrieval. In the next chapter we investigate ways to improve both the precision and recall of image retrieval by combining both modalities.

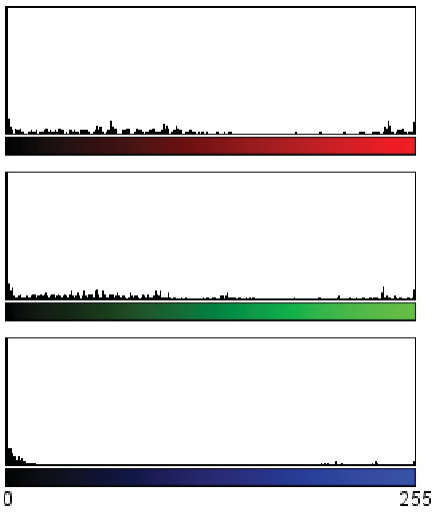


Figure 19: Histogram of the Median Image of Figure 12 in IHS Model

Figure 20: Partitioning the Image for Visual Retrieval

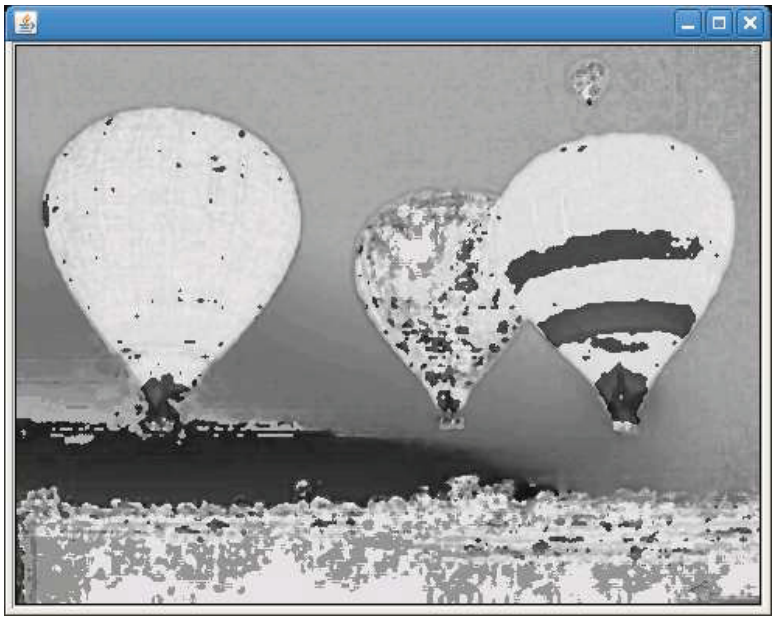

Figure 21: Grey Image of Figure 12

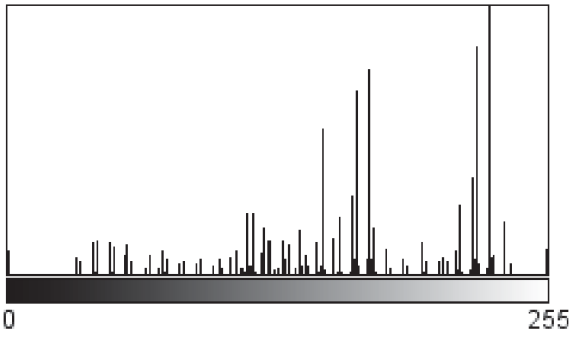


Figure 22: Grey Histogram of the Image of Figure 12

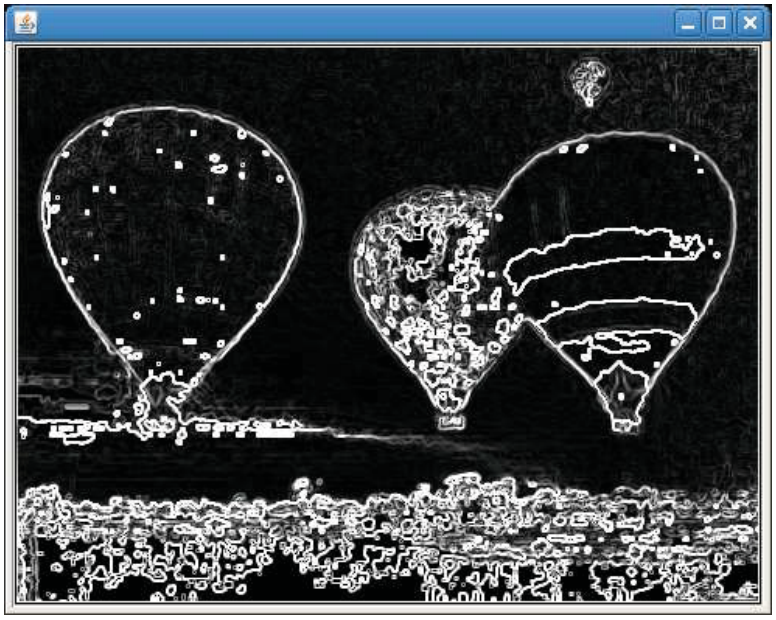

Figure 23: Gradient Image of the Image of Figure 12

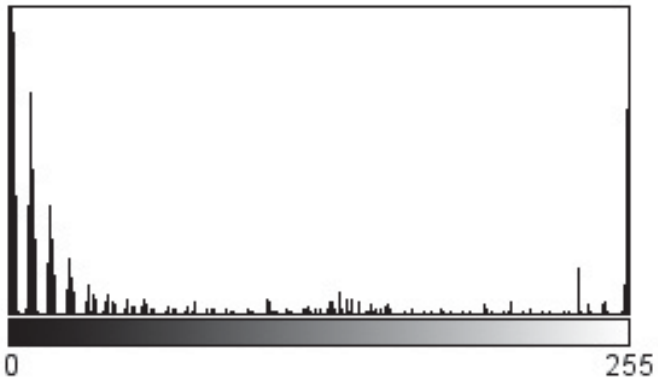


Figure 24: Gradient Magnitude Histogram of the Image of Figure 12

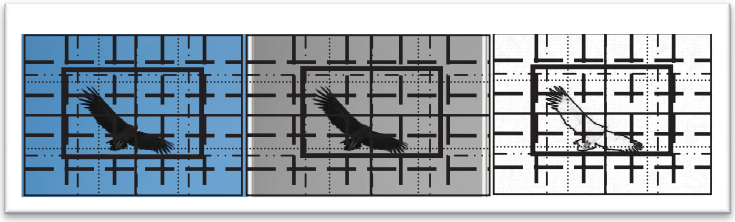

Figure 25: Block-based Visual Features Extracted from Color, Grey-Scale and Gradient-Magnitude Images

Table 12: Visual Feature Vector

| Description | Region | Histogram |
|---|---|---|
| 1-36 | 2X2 division | IHS |
| 37-117 | 3X3 division | IHS |
| 118-261 | 4X4 division | IHS |
| 262-486 | 5x5 division | IHS |
| 487-495 | Center | IHS |
| 496-504 | Thumbnail | IHS |
| 505-516 | 2X2 division | Gradient Magnitude |
| 517-543 | 3X3 division | Gradient Magnitude |
| 544-591 | 4X4 division | Gradient Magnitude |
| 592-666 | 5X5 division | Gradient Magnitude |
| 667-675 | Whole Image | IHS |
| 675-677 | Whole Image | Grey Scale |
| 678-680 | Center | Grey Scale |
| 681-683 | Whole Image | Gradient Magnitude |
| 684-686 | Center | Gradient Magnitude |

Table 13: Impact of the Different Visual Features on the IAPR TC-12 Collection

| Description | MAP | P10 | P20 | P30 | Recall |
|---|---|---|---|---|---|
| Removing Color Distance | 0.0611 | 0.2564 | 0.1705 | 0.1308 | 601 |
| Removing All Color | 0.0343 | 0.1718 | 0.1115 | 0.0829 | 432 |
| Removing Gradient Similarity | 0.0620 | 0.2513 | 0.1718 | 0.1316 | 619 |

# Chapter 5

# Fusion Methods

The idea behind combining text and visual features in image retrieval is to make use of their diverse nature to find results that otherwise could not be retrieved by the single-medium retrieval methods. There are various ways to combine these features. They can be categorized according to the phase of the combination in relationship to the retrieval step, as well as the data that the fusion is applied to. For the combination phase, the fusion between visual and textual methods can happen early or late in the retrieval process. The following distinctions can be made :

- Early fusion methods
- Late fusion methods
- Mixed methods

Early fusion methods are applied before the actual matching of the query to the collection. Late fusion methods use either the results returned from different retrieval models or the scores of the documents returned by these models to build an improved list of results. On the other hand, mixed methods are applied during any of the retrieval steps.

In terms of the data categories that the fusion is applied to, these can be divided into the following categories or any combination of them:

- The data collection
- The queries
- The results

When applied to the data collection, early fusion methods can be used to build separate indices, or a combined index of the visual and text features. Early fusion of queries necessitates the availability of both a text query and a visual query beforehand. The methods adopted in this dissertation belong to the mixed fusion category, where an initial visual query is processed, the results are used to expand the text query, then a new text query is formulated and processed. Hence, the actual fusion is between the annotation from the data collection and the text query.

Recall from Section 4.1 that the experiments conducted on the IAPR TC-12 collection for the 2007 ImageCLEF benchmark revealed an improvement in the results obtained by the run combining both visual and textual retrieval systems over those using a single modality in a majority of the topics (See run 1 of Table 8 - Section 4.1.6). The fusion method used in the 2007 benchmark (Section 4.1) was a simple combination of the results from the text and visual retrieval engines, ranking higher their intersection, followed by the text retrieval results, then the visual retrieval results.

In the next section, a more elaborate approach for expanding the text query with terms from the top results of the visual query is presented. This is the Pseudo-Relevance Feedback (PRF) component of our method. Building on this approach, Section 5.2 proposes a robust method for filtering and weighting the terms used for expansion, the Semantic Filtering component (SF), followed by Section 5.4 which compares the different text retrieval models when used with the PRF-SF method. Section 5.5 demonstrates the complementarity of the visual and textual retrieval components used. Section 5.6 revisits the diversity factor, comparing the diversity of results returned by text-only retrieval against those of the PRF-SF method, and Section 5.7 aims to render a transparent view of the PRF-SF method by providing concrete examples of queries on which the method improved the results over the single-modal retrieval methods, and others where it achieved less satisfactory results.

## 5.1 Fusion of the Results Using Simple Query Expansion (PRF)

Given the low performance of image retrieval algorithms, as well as the relative maturity of text retrieval methods compared to them, we wanted to investigate if the combination of two low-performing algorithms with pseudo-relevance feedback can result in much better performance. We thus developed a hybrid system and participated in the 2008 ImageCLEFPhoto task. The main purpose of the experiments was to maximize the Mean Average Precision (MAP) of the results. In order to evaluate the method, we conducted experiments on the track's collection of 20,000 tourist photographs [Arni *et al.*, 2008 printed in 2009]. As presented in Section 3.3.2, the collection consists of equal-size mostly color photographs taken in various locations around the world.

To combine the results from the two media searches, the confidence level in the visual results (i.e. the level of proximity from the query images) was taken into consideration. A maximum of three highest ranked images is taken from the visual query results depending on the confidence score, followed by the text results after query expansion. As we can see from Table 14, this simple re-ranking method only improved a little on the run that utilized only pseudo-relevance feedback in the official results (run clacIRTX vs. run clacTxNr) . When adding the *description* field, it lowered considerably the precision. Supplementary fusion methods could be useful on top of pseudo-relevance feedback in case of the availability of little textual data or text retrieval with low-precision.

Several ways of query expansion were experimented with:

- Method 1: The highest ranked *n* results from the text search engine were passed as additional example images to the visual search. Values of *n* from 1-5 were experimented with.
- Method 2: The highest ranked text search results were used to expand the text query.
- Method 3: Noun synonyms from WordNet were added to the query.
- Method 4: All terms in the annotations of the highest ranked visual results were added to the text query.

The last method was the only one found to be beneficial in improving the MAP of the results, and is the only one reported in this thesis.

Six runs experimenting with the block-based visual retrieval as well as with query expansion were submitted to ImageCLEF 2008. Table 14 shows the results, published in [El Demerdash *et al.*, 2009b], that we obtained in comparison to the mean, median and best runs of the track, taken from the best four runs from each participating group (25 groups and 100 runs in total). As Table 14 shows, despite the poor performance of the visual (clacIR - Map=0.0552) and text retrieval components (clacNoQE - Map=0.2034), better results can be obtained through pseudo-relevance feedback and the inter-media fusion of the results (clacIRTX - Map=0.2622). As expected, the highest Mean Average Precision (MAP) was obtained by the runs that utilized the maximum resources and methods combining both visual and text retrieval. Despite the weak results of the visual-only run (clacIR), the block-based method used (presented in Section 4.2.2) was appropriate for the top retrieved results. The low MAP score of the visual-only run is due to the simple features chosen as a conscious trade-off between precision and execution time. The run that obtained the best score in the experiments submitted to ImageCLEF2008 is the one using pseudo-relevance feedback for query expansion according to the last method listed above, (clacIRTX, and with the Description field clacDesc - Map=0.3419). While the MAP of the visual only run (clacIR) is only 0.055, its precision at five retrieved documents (p5=0.328) is significantly higher than that of the text only run clacTX (p5=0.236). For this reason, the highest ranked document was used for the expansion of the text query. This was only done if the document meets a confidence level determined empirically. The confidence score is assigned based on the proximity score to the query image.

Figure 26 shows the breakdown of the *MAP* by topic for five of the official six runs submitted sorted by the precision per topic of the best run (clacIRTX). It is notable that the runs with feedback, (clacTxNr) and (clacIRTX), performed consistently better than the single media runs, (clacTX) and (clacIR), as well as the combined run without feedback (clacNoQEMX), except in cases where there was a significant divergence between the visual and text search results. The relevance feedback mechanism tends to average between these diverging results.

As illustrated on the ImageCLEFPhoto 2008 data, the use of simple, light-weight, low-cost and relatively lower-precision retrieval systems can be significantly improved through the use of pseudo-relevance feedback.

Table 14: Results at ImageCLEFPhoto 2008 Using PRF from the Top Visual Result Only

| Run ID | Modality | **MAP** | P10 | P20 | P30 | GMAP | Rel | F-measure |
|---|---|---|---|---|---|---|---|---|
| clacTX | Text | 0.1201 | 0.1872 | 0.1487 | 0.1462 | 0.019 | 1155 | 0.1741 |
| clacTxNr | Text | 0.2577 | 0.4103 | 0.3449 | 0.3085 | 0.1081 | 1859 | 0.3290 |
| clacIR | Visual | 0.0552 | 0.2282 | 0.1615 | 0.1214 | 0.0268 | 0629 | 0.1877 |
| clacIRTX | Mixed | 0.2622 | 0.4359 | 0.3744 | 0.3308 | 0.1551 | 1630 | 0.3546 |
| clacNoQE | Mixed | 0.2034 | 0.3205 | 0.2705 | 0.2487 | 0.0780 | 1701 | 0.2875 |
| clacNoQEMX | Mixed | 0.2180 | 0.4026 | 0.3269 | 0.2855 | 0.1290 | 1546 | 0.3384 |
| clacDesc | Mixed | 0.3419 | 0.5051 | 0.4256 | 0.3726 | 0.1794 | 2401 | |
| ImageCLEF best/team | | | | | | | | |
| Average run | N/A | 0.2187 | | 0.3203 | | | | |
| Median run | N/A | 0.2096 | | 0.3203 | | | | |
| Best run(Manual) | N/A | 0.4288 | | 0.6962 | | | | |
| Best run(Automatic) | N/A | 0.4105 | | 0.5731 | | | | |

There was little correlation between the visual descriptors chosen and the annotations of the images. It is possible in this case to have confidence in the top results only. A visual search system based on supervised training methods would likely have a much higher correlation. While this leads to higher precision, more overlap with the text results would render pseudo-relevance feedback less useful.

## 5.2 Pseudo-Relevance Feedback with Semantic Restrictions (PRF-SF)

Semantic query expansion is a method that has often been tackled unsuccessfully in the Information Retrieval domain [Voorhees, 1994]. The apparent reason for this is the introduction of a too-high ratio of noisy non-relevant terms to actual relevant terms in the expanded query. For this reason, a more prudent approach involving the semantic filtering of the expansion has more potential than the direct expansion of the query. Semantic filtering is the removal from query expansion of terms not related in meaning to the original terms in the query.

Figure 27 illustrates the architecture of the proposed PRF-SF method. The method is capable of tackling initial queries in the form of text only or text with example images. In the case of a text-only query an initial text retrieval step can be performed and the best retrieved results used as example images. The visual queries are processed by the content-based engine.

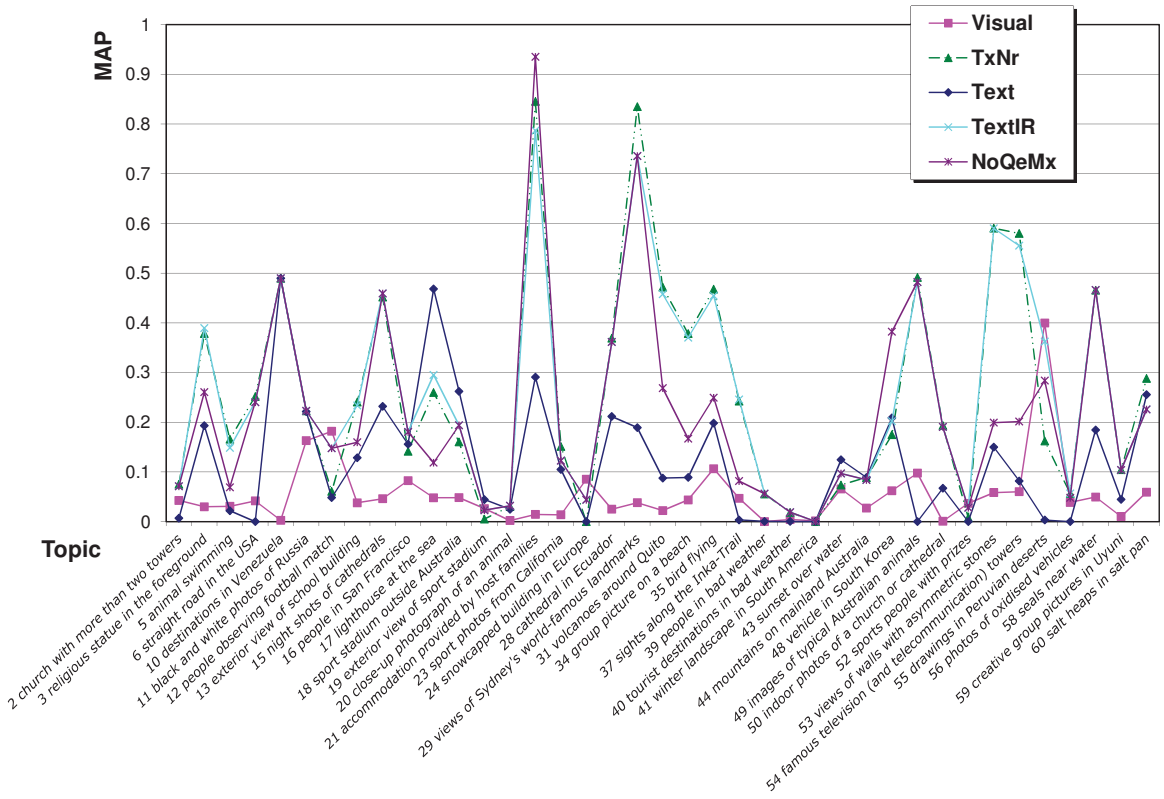


Figure 26: Comparison of Runs by MAP of Each Topic on the ImageCLEFPhoto 2008 Queries

Common ways for text query expansion include adding synonyms and other semantically related terms to the query. However, as described earlier, according to the experiments on the IAPR TC-12 collection, this approach may lead to the introduction of too many noisy terms. Instead, the extraction of related terms from the highest-ranked results retrieved by the content-based system described in Section 4.2 is a more effective alternative. Semantically constrained query expansion was attempted on the 2008 IAPR TC-12 collection as well as the 2009 Belga ImageCLEFPhoto task data. The official results on the Belga dataset were published in the conference working notes [El Demerdash *et al.*, 2009a] and the proceedings

of CLEF [Demerdash *et al.*, 2009], while the IAPR dataset results are published in the IEEE MMSP 2009 proceedings [El Demerdash *et al.*, 2009c].

In 2008, the ImageCLEFPhoto queries consisted of a query topic, a narrative describing the exact relevance criteria and three unlabeled example images as visual queries. There were 39 queries in the 2008 ImageCLEFPhoto collection. An example of a query topic is *bird flying*, while its narrative states that *Relevant images will show one or many birds in the act of flying. Birds that are not flying are not relevant. Other flying objects that are not birds are not relevant either.* The three example images were of the same size, but not necessarily the same orientation. They were not included in the data set, and hence were not returned in the top five images used for query expansion.

For the data sets used in the experiments on the 2009 benchmark, all the terms associated with the image are extracted except for stop words. Due to the much larger size of the data set (approximately 500,000 images) compared to the IAPR TC-12 collection (20,000 images) used in previous years, as well as the scarcity of computing resources at the time, we resorted to reducing the index by eliminating some of the descriptors we used previously, such as the grey-level and gradient-magnitude descriptors.

In the expansion phase, the query is expanded with terms potentially related to the query (see Figure 27). In order to expand the query without introducing noise, the candidate text is compared to the query topic for potential semantic similarity. If the image is found to be potentially related to the topic, the text query is expanded with the relevant terms. To assess the possibility of a relationship between a given image's annotation and a topic in question, the minimum threshold of one common non-grammatical word (i.e. non stop-word) is used, due to data sparseness.

The top $n$ results of the visual engine are exploited for query expansion. As seen in Figure 27, the system first extracts the terms related to the candidate. The extracted terms are then passed through a filter constructed by the semantic expansion of the text query. The purpose of the filter is to find a minimum common denominator between the topic of the query and the potential expansion image. It also serves to filter out contradictory and irrelevant terms from query expansion to avoid introducing noise. Consequently, the text query is expanded with terms extracted from images with common visual and semantic similarities.

In a final phase, the results from the text and visual queries are post-fused through a re-ranking mechanism to increase recall and ensure diversity and coverage.

The purpose of the query expansion module is not only to augment the query by adding new candidate terms related to it, but also to enhance it by adding weights to its key terms and filtering out potentially noisy terms from expansion. An example is the query *flying bird*. A top visual match is annotated with *condor flying*, hence, the matching term *flying* is given more weight. This approach results in stressing that the bird be flying through redundancy.

Another important step is to avoid expanding the query with named entities that do not have a semantic relationship with the query. This is crucial in photographic collections, since by their nature, photographs and image queries are often bound by geographical constraints. For example, a query requesting *straight road in the USA* has the *USA* as a geographical constraint. In order to ensure that potential expansion images do not introduce conflicting geographical terms in the query (i.e. locations outside the USA), a filter is first built from the location specified in the query. This feature makes use of WordNet [Fellbaum, 1998], a lexical database, by traversing its *PartMeronym* hierarchy. A *PartMeronym* is a relationship between two nouns where the child noun constitutes a part of the parent noun (e.g. engine-car). For geographical locations, this translates by the divisions of the parent noun. For example for the USA, a traversal of the hierarchy produces the names of the states, then major cities and towns followed by specific locations. While similar filters are possible for common nouns and using other relations such as Hyponymy (sub-classes of a term - e.g. dog-animal), the expansion was limited to named-entities, so as to avoid the problem of disambiguation of the specific sense of the term. This problem will be dealt with in Section 5.3.

The text query expansion module involves a pseudo-relevance feedback mechanism and the fusion of the text and visual search results. First the visual query is executed, then the highest results obtained are used to expand the text query. An additional fusion is performed on the results obtained from both engines.

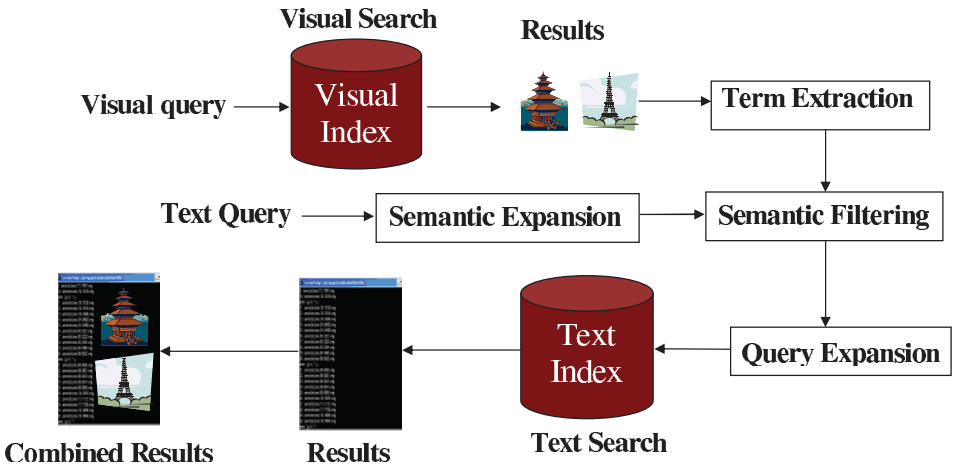


Figure 27: Overview of the Fusion Method

### 5.2.1 Results on the IAPR TC-12 Dataset

Evaluation of results is performed using the same version of TREC-EVAL used in the official ImageCLEF-Photo track. Table 10 shows the main precision and recall measures: precision at 5, 20, 30, 100 documents retrieved (P5, P20, P30, P100 respectively), Mean Average Precision (MAP) and Recall for the text-only, visual-only, and combined method, as well as the automatic run with the highest MAP at ImageCLEFPhoto 2008 and the best individual score per category from all ImageCLEFPhoto runs combined. The total number of relevant documents for all 39 queries is 2401.

Table 15: Results on ImageCLEFPhoto 2008 Data

| Description | P5 | P20 | P30 | P100 | MAP | Recall |
|---|---|---|---|---|---|---|
| Visual | 0.387 | 0.178 | 0.132 | 0.062 | 0.064 | 0.259 |
| Text | 0.472 | 0.383 | 0.330 | 0.210 | 0.302 | 0.754 |
| **PRF-SF** | **0.780** | **0.654** | **0.583** | **0.334** | **0.505** | **0.859** |
| Best MAP | 0.723 | 0.573 | 0.486 | 0.283 | 0.411 | 0.790 |
| Best score | 0.728 | 0.573 | 0.488 | 0.285 | 0.411 | 0.842 |

As Table 15 shows, the proposed PRF-SF method, combining content-based and text retrieval using auto-relevance feedback with semantic filtering, outperforms the highest precision and recall measures for an automatic run, as well as the best individual score per category obtained at ImageCLEFPhoto 2008. The combined results also demonstrate an increase in MAP over the text-only retrieval of about 67.5%. The bulk of this significant gain can be attributed to the introduction of new relevant terms through the controlled query

expansion process. These terms are often semantically related to the query topic, such as concrete examples of a concept (*bird-condor*), a relevant instance (*church with more than two towers-St. Patrick's Cathedral, Melbourne*), a geographical sub-region (*USA-Colorado*), or a synonym (*straight road-highway*). The rest of the improvement is a result of the use of redundancy to stress key terms of the query. It is also worth noting that post-fusion of the text and visual results had an insignificant contribution to MAP and about 5% increase in retrieved documents.

The block-based method achieves 6.4% MAP and 38.7% precision at 5. In comparison, experiments conducted with the MPEG-7 descriptors ScalableColor, ColorLayout and EdgeHistogram combined, yielded a lower MAP of 2.9% and lower precision at 5 of 12.67%. The block-based method also yields almost 100% increase in the number of retrieved documents. As with most content-based systems (Section 2.2), the most significant feature contribution comes from the color histograms (Section 4.2.7). The histograms of the gradient magnitude image constitute a boosting factor, while the grey-level histograms are not an essential factor. As for the regional divisions, they are indispensable for a higher early precision, especially for the color histograms.

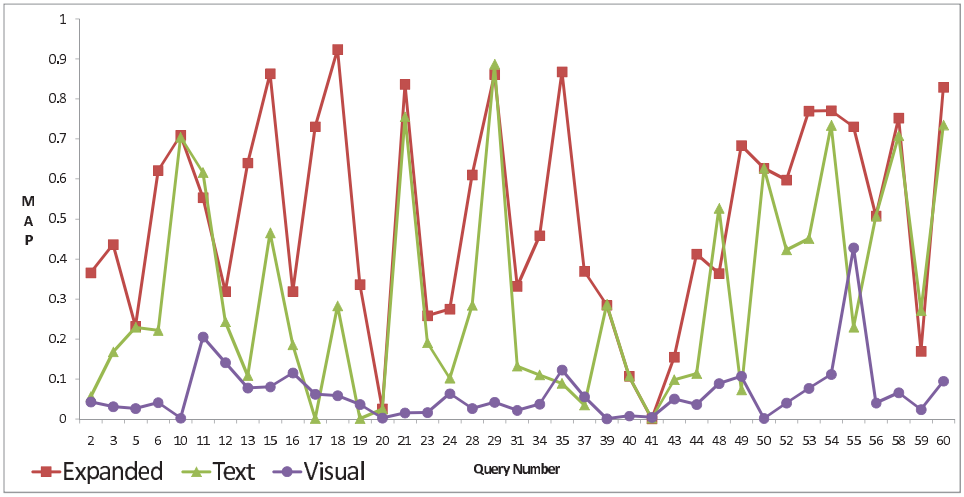


Figure 28: Performance (MAP) per Topic of the Visual, Text and Combined Retrieval

Figure 28 shows the individual topic MAP performance by the visual, text and combined retrieval using auto-relevance feedback (note that query numbers are not in series). Visual retrieval significantly outperforms text retrieval in four queries. These queries greatly improved through expanded retrieval. Seven queries did

not undergo any expansion mostly due to their poor content-based results. Only three out of the remaining 32 queries lost tangible precision due to noisy query expansion and/or faulty weight increase. About 25 queries improved substantially after the expansion, the majority of which would not have adequate answers without the expansion process. An example is query 17 *lighthouse near the sea*, which was expanded with the specific names of two lighthouses *Cape Otway* in Australia and *Ushuaia* in Argentina, while assigning more weight to both terms *lighthouse* and *sea*.

### 5.2.2 Results on the Belga Set

Table 16 shows the results on the ImageCLEFPhoto 2009 Belga data sets. The first two runs are purely visual and textual respectively. The *PRF* run combines visual and text retrieval using the Pseudo-relevance feedback mechanism described in Section 5.2 and separate queries for each cluster, the results of which are then combined using a simple interleaving method, taking one result in turn from the top of the list of results for each query. The *Combined* run uses the same method as the *PRF*, while combining all clusters information into one query. *P10* and *P20* are the Precision figures at 10 and 20 retrieved documents respectively. *CR10* and *CR20* are the Cluster Recall levels at 10 and 20 retrieved documents, while the F-measure reported in these tables employs P10 and CR10 similar to the official F-measure used at the 2009 ImageCLEF campaign.

Table 16: Results on ImageCLEFPhoto 2009 Queries (Belga Dataset).

| **Description** | **P10** | **P20** | **CR10** | **CR20** | **MAP** | **Rel_Ret** | **F-measure** |
|---|---|---|---|---|---|---|---|
| Visual | 0.0960 | 0.0990 | 0.2980 | 0.4340 | 0.0060 | 657 | 0.1452 |
| Text | **0.7540** | **0.7800** | 0.6877 | 0.7525 | **0.4879** | **19148** | **0.7193** |
| With PRF | 0.5820 | 0.6770 | **0.7334** | **0.8482** | 0.4221 | 17880 | 0.6490 |
| Combined Clusters | 0.6200 | 0.7090 | 0.6822 | 0.7972 | 0.4531 | 18387 | 0.6496 |

The results demonstrate that using text only queries outperforms the pseudo-relevance feedback runs in the F-measure (0.7193) as well as precision (0.754) and (0.78). However, the diversity of the pseudo-relevance feedback runs tends to be higher. The visual-only run rated very poorly. Indeed, the successful pseudo-relevance appeared to stem from expanding using the text associated with the example images, which were eliminated from the gold standard and did not count as valid results.

Tables 17 and 18 show the breakdown of these runs by query set (queries where the cluster information

was given and queries without cluster information respectively).

Table 17: Queries with Given Clusters (Belga Dataset).

| Description | P10 | P20 | CR10 | CR20 | MAP | Rel_Ret | F-measure |
|---|---|---|---|---|---|---|---|
| Visual | 0.0720 | 0.0820 | 0.2603 | 0.3934 | 0.0026 | 241 | 0.1128 |
| Text | **0.7400** | **0.7660** | **0.7796** | **0.8693** | **0.4595** | 8778 | **0.7593** |
| With PRF | 0.5400 | 0.6900 | 0.7562 | 0.8772 | 0.4207 | 8664 | 0.6300 |
| Combined Clusters | 0.6000 | 0.7220 | 0.6741 | 0.7702 | 0.4476 | **8793** | 0.6349 |

Table 18: Queries without Given Clusters (Belga Dataset).

| Description | P10 | P20 | CR10 | CR20 | MAP | Rel_Ret | F-measure |
|---|---|---|---|---|---|---|---|
| Visual | 0.1200 | 0.1160 | 0.3357 | 0.4757 | 0.0095 | 416 | 0.1768 |
| Text | **0.7680** | **0.7940** | 0.5958 | 0.6358 | **0.5164** | 10370 | **0.6710** |
| With PRF | 0.6240 | 0.6640 | **0.7106** | 0.8192 | 0.4234 | 9216 | 0.6645 |
| Combined Clusters | 0.5680 | 0.6200 | 0.6902 | **0.8242** | 0.4585 | **9594** | 0.6641 |

There is a significant difference between the precision and cluster recall at ten (P10 & CR10) and at 20 (P20 & CR20) retrieved results. Unexpectedly, precision increases with retrieved results ($P20 > P10$), and up to the top hundred results (P100). This is due to some noisy early results introduced by the errors in visual retrieval. Contrary to ImageCLEFPhoto 2008 the F-measure was computed that year using a cut-off of the first ten results, which was a disadvantage to this method. The MAP and the Relevant Retrieved figures are promising and show consistency over the different topics.

Figure 29 shows the individual queries MAP performance of each of the four runs, while Figure 30 shows the Cluster Recall at 10 retrieved results of the three textual and mixed runs. We note that the text-only run shows a higher standard deviation than the pseudo-relevance feedback method, especially due to the very low precision of two queries (Queries 10 and 43). In both cases the PRF method managed to reasonably answer the queries due to the visual input. Combining the cluster information in one query improves precision but decreases cluster recall.

The experiments at ImageCLEF 2009 with applying semantic selectional restrictions aimed to enhance cross-media pseudo-relevance feedback and attempt different methods of query formulation for clustered queries. The findings show that in the presence of valid results from a visual retrieval system, pseudo-relevance feedback can be successfully implemented and enhances the diversity of the results; however, the precision of the text only retrieval is still better.

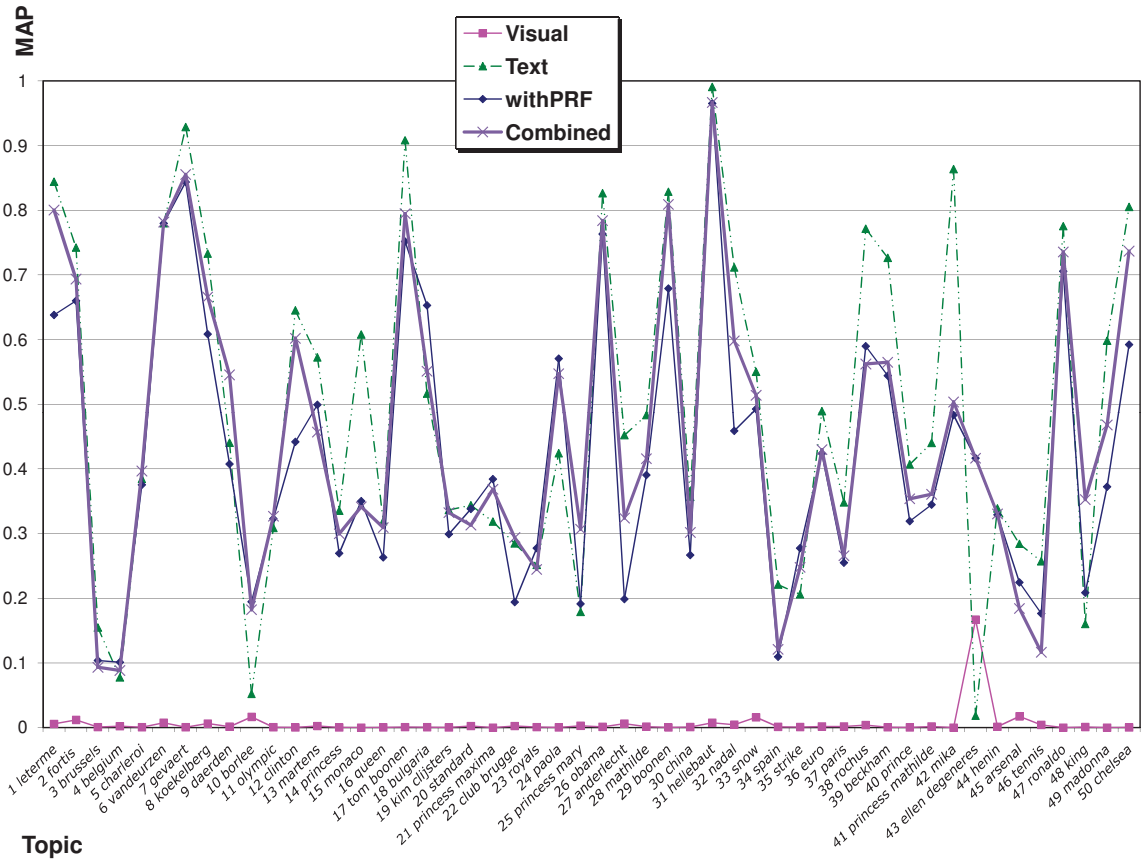


Figure 29: Map by Query.

Two factors can shed a light on the reason the filtered semantic expansion achieved a lower MAP than the text-only method on the Belga set as demonstrated in Section 5.2.2: The first is noise introduced through using a corpus-specific stop-word "Belga" for the expansion of almost all queries. As is customary in news agencies' press releases, annotations of the Belga set started with the term "Belga" as the source of the image. This provided a false related term for the expansion module, and hence unrelated images were used in the expansion. The second factor is the deterioration of the visual results returned at the top of the content-based engine with the much larger dataset. In order to robustly remedy these two problems, a few collection-specific stop-words need to be added to the stop-words list, and a more restringing method for filtering the expansion terms and weighting them is necessary.

In the next section, we describe a more robust enhanced semantic expansion filtering mechanism, as well

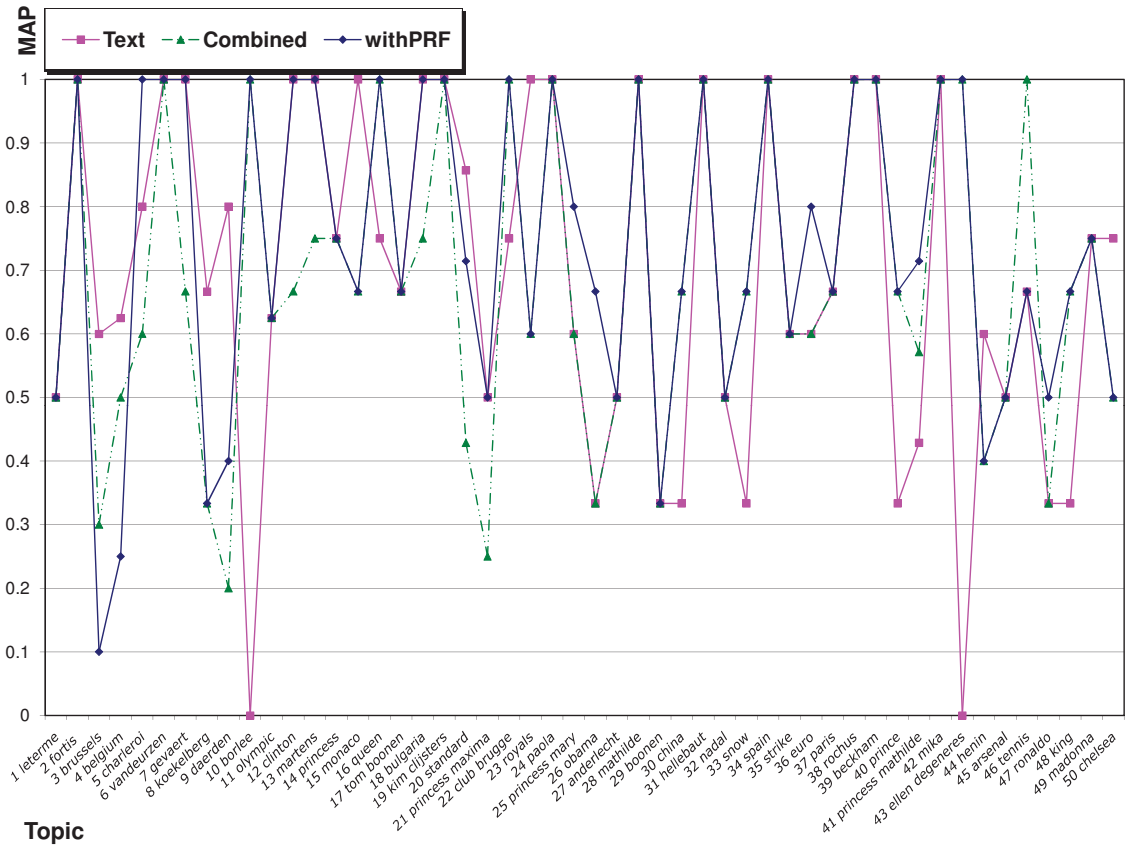


Figure 30: Cluster Recall by Query on the Belga Data Set.

as the results from running it on the datasets.

The method proposed in this dissertation leverages the text retrieval results through the incorporation of image retrieval, thus making use of the best results returned by the content-based process and the overall higher precision and recall of text retrieval. A similar method has been investigated in [Maillot *et al.*, 2006]; however, the method presented here introduces the notion of semantic filtering of the expansion, which render better and more reliable results. In [Maillot *et al.*, 2006], the terms from the top three results are used for query expansion. This leads to a significant improvement of the precision (Map=0.3337 for query expansion and 0.1619 for text-only retrieval). The reason for using the top three images only is a decreasing confidence in the visual results. However, when we employ enhanced semantic filtering involving different semantic relations as will be demonstrated in section 5.3, we are able to use the top 20 results, without introducing significant noise, since we are confident that the terms introduced are semantically related to the original

query.

## 5.3 Enhanced Semantic Filtering

The PRF-SF method we developed consists in confining the visual expansion using a visual similarity threshold, instead of a semantic similarity, in addition to creating a related list against which the potential expansion terms are compared for weights. Using the IAPR TC-12 collection as a development set, the visual expansion limit was empirically determined to be within about 7% of the maximum theoretical visual distance of the 686 features used. Only images within this threshold are potential expansion candidates, instead of the one-common-term semantic threshold applied to the annotation text.

As for the related list constructed from the query terms, it contains terms from the following relations extracted from the WordNet database:

- Hyponyms: Terms having a type-of relationship with the query term. For example for the terms “road”, “driveway”, and “highway” are hyponyms.
- Part Meronyms: Terms representing a part of a whole. For example a “bend” or a “curve” are part meronyms of “road”.

Other relations were investigated but found to not bear a significant effect on the results. These include:

- Instance Hyponyms: A specific named instance of a hyponym as defined above. For example, “Champs Elysees” is an Instance Hyponym of “road”.
- Member Holonyms: Terms with which the term in question have a part-of relationship. For example a “bird” is a part of a “flock” or “aves”.

After these terms are gathered in a list, the expansion terms extracted from the annotation of the highest scoring images matching any of the example images, and that meet the minimum visual distance threshold, are compared against the related-list. If a term is found on the list, it is added to the query with a full weight, otherwise the term is added to the query with a partial weight (less than 1).

An example of an expansion is the query "Obama", which was expanded with the following word stems: michell, wife, democrat, presidenti, candid, u s, senat, barack, democrat, illinoi, wave, introduc, ralli, univers, illinoi, chicago, pavilion, midst, offici, campaign, trip, iowa, new hampshir, formal, announc, candidaci, epa, tannen, mauri. Section 5.7 demonstrates and analyzes examples of successful and unsuccessful query expansions.

As mentioned earlier, we developed this approach on the IAPR dataset, and tested it on the Belga set. The results on the Belga set after applying these enhancements are presented in Table 19. They demonstrate a significant improvement in the pseudo-relevance feedback method when employed with semantic filtering over the text-only retrieval in both MAP and recall. Applying the Wilcoxon signed-rank test, this result is confirmed statistically significant with very high confidence (z=3.65, p=0.0001), as well as a significant improvement over the result when employing PRF without semantic filtering (z=4.11, p$<$0.0001)[1]. We also note that the use of PRF alone without filtering results in a major deterioration of all measures, even compared to the text-only results.

Table 19: Results on the Belga Set Using the Filtering Method.

| **Description** | **P10** | **P20** | **MAP** | **Recall** |
|---|---|---|---|---|
| Visual | 0.0960 | 0.0990 | 0.0060 | 657 |
| Text | 0.7540 | 0.7800 | 0.4879 | 19148 |
| PRF without SF | 0.4580 | 0.4490 | 0.2227 | 11810 |
| PRF-SF | 0.6940 | 0.7440 | 0.5300 | 20407 |

## 5.4 Different Retrieval Models Compared With Fusion

For the sake of completeness, a comparison between the vector-based TF-IDF model and the probabilistic BM25 and PL2 methods on both datasets, similar to that in Section 4.1.4, is presented in this section in order to investigate the effect of different text retrieval models within the context of multi-modal retrieval methods. Once again, the difference was found to be minimal. Table 20 shows the results on the IAPR TC-12 dataset, table 21 on the Belga dataset, and Table 22 on the Belga set when employing a post-fusion retrieval mechanism.

[1] tests conducted using the online tool at http://vassarstats.net/wilcoxon.html

Table 20: Comparison between TF-IDF and Probabilistic Models on the IAPR TC-12 Data Mixed Retrieval

| Model | MAP | P10 | P20 | P30 | Recall |
|---|---|---|---|---|---|
| TF-IDF | 0.5043 | 0.7410 | 0.6397 | 0.5709 | 2082 |
| BM25 | 0.4996 | 0.7359 | 0.6449 | 0.5701 | 2086 |
| PL2 | 0.5024 | 0.7462 | 0.6423 | 0.5718 | 2048 |

Table 21: Comparison between TF-IDF and Probabilistic Models on the Belga Mixed Retrieval

| Model | MAP | P10 | P20 | P30 | Recall |
|---|---|---|---|---|---|
| TF-IDF | 0.5335 | 0.7260s | 0.7510 | 0.7713 | 20443 |
| BM25 | 0.5339 | 0.7160 | 0.7460 | 0.7673 | 20464 |
| PL2 | 0.5321 | 0.7360 | 0.7490 | 0.7627 | 20508 |

## 5.5 Complementarity of the Text and Visual Retrieval Components

In order to demonstrate the complementarity of the visual and textual retrieval components, Figures 31, 32, and 33 show the results per topic on the 39 queries used for the ImageCLEFPhoto 2008 campaign for different metrics using all retrieval methods. The first two of these figures present the single-modality results, and the last one shows the mixed-modality results.

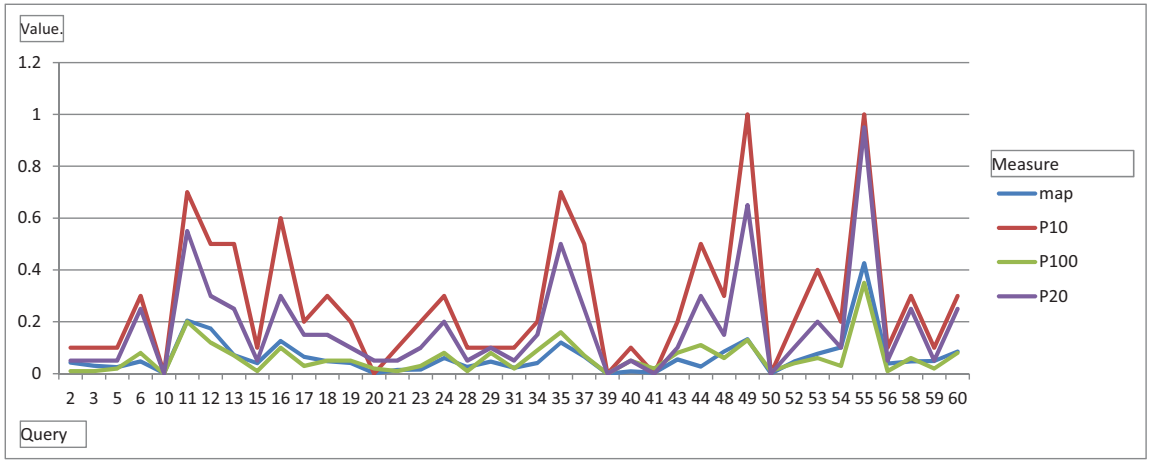


Figure 31: Visual-only Retrieval Metrics on IAPR TC-12

It can be deduced from these figures that the visual retrieval method often achieved a much better level

Table 22: Comparison between TF-IDF and Probabilistic Models on the Belga Mixed Retrieval with Post-Retrieval Fusion

| Model | MAP | P10 | P20 | P30 | Recall |
|---|---|---|---|---|---|
| TF-IDF | 0.5856 | 0.7680 | 0.7870 | 0.7847 | 21301 |
| BM25 | 0.5866 | 0.7680 | 0.7870 | 0.7867 | 21326 |
| PL2 | 0.5854 | 0.7780 | 0.7790 | 0.7840 | 21376 |

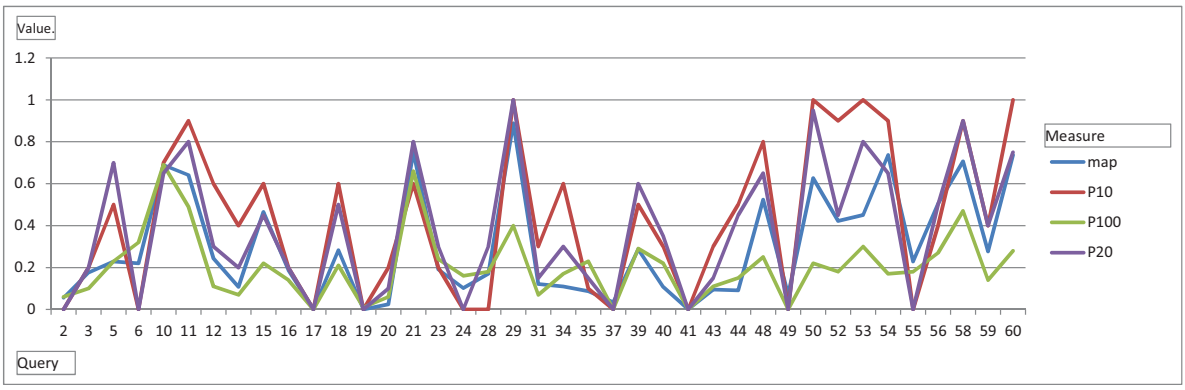


Figure 32: Text-only Retrieval Metrics on IAPR TC-12

of precision in the first 20 retrieved documents (P20) in those instances where the text retrieval component failed to produce satisfactory results (for example topics 2, 6, 17, 19, 24, 28 and 37).

Figure 34 shows the detailed performance by topic of the visual and text systems as well as their combination. In the few cases where the text retrieval obtained a higher MAP, the combined result was affected by noise induced from the visual results. On the other hand, the visual results achieved higher precision in some topics because of the reliance mainly on the text results, due to the higher confidence in them.

These observations on the development set provide the premise for expanding the queries based on the top visual retrieval results with in a threshold of similarity, combined with semantic filtering.

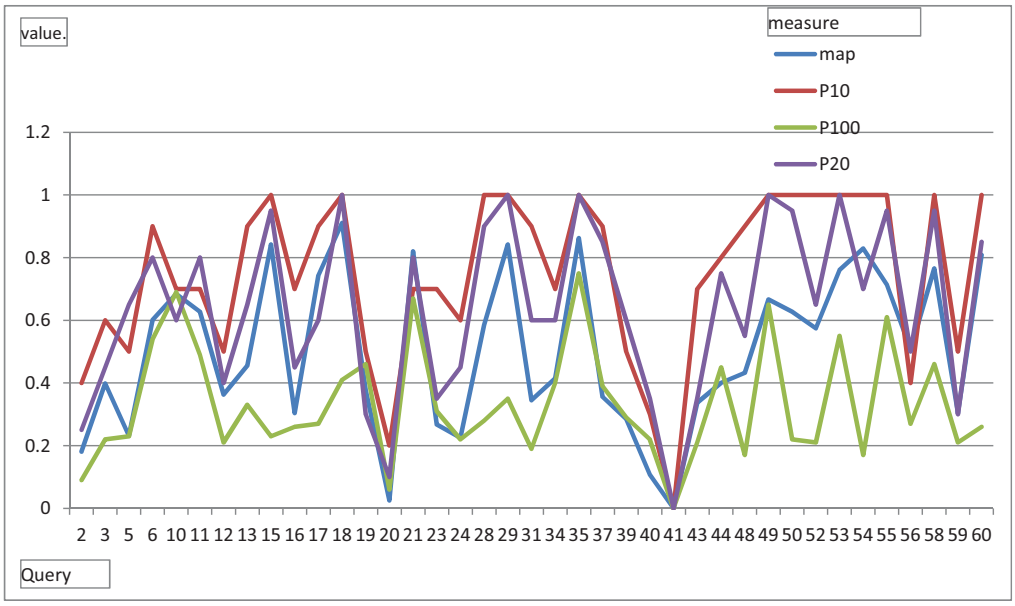


Figure 33: Combined Text and Visual Retrieval Metrics on IAPR TC-12 Using the PRF-SF Method

## 5.6 The Diversity Factor

As discussed in Section 1.5, an important drawback of text-only retrieval is the lack of diversity in the results. This is especially true when one sense of a query term occurs significantly more than the other senses, since text retrieval hinges mostly on statistical frequency measures. Table 23 compares using the diversity metric employed by the ImageCLEFPhoto campaign Cluster Recall (CR). The most reflective of the cluster recall measure are those at 20 and 30 recalled documents (CR@20 and CR@30) since this level roughly corresponds to the first page of results viewable on a user interface.

From this table, we note that the PRF-SF method improves the cluster recall figures at all levels, but more so at the more significant levels of the top 20 and 30 recalled documents (0.6645 for text retrieval CR@20 vs. 0.7205 for the query expansion method, and 0.7078 vs. 0.7918 for CR@30). This can be attributed to the visual features employed in conjunction with the text query, thus diversifying the results. To revisit the problem illustrated in Section 1.5, the combination of text and visual features is more capable of alleviating the problem of lack of diversity in the results than the use of text-only methods. Furthermore this method is more suited for accommodating relevance feedback from the user than text retrieval methods.

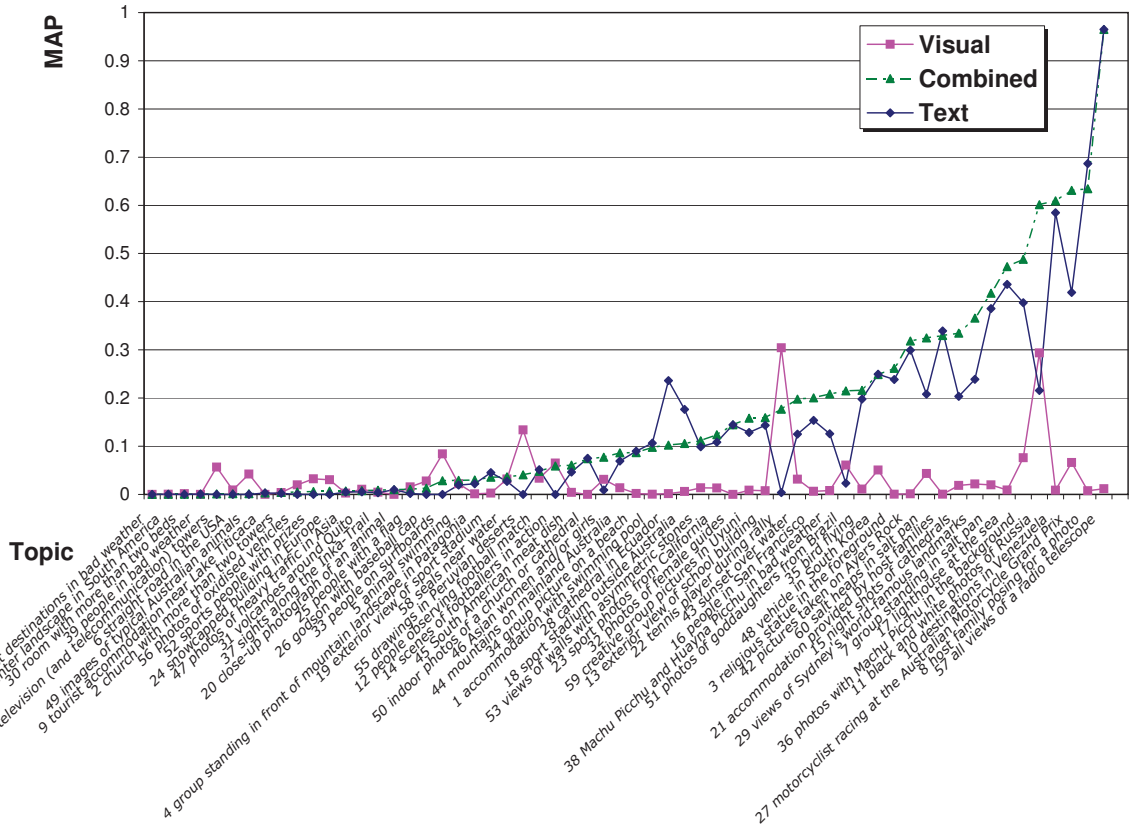


Figure 34: Comparison between Text, Visual and Combined Results by Topic on the 2007 ImageCLEFPhoto Queries (IAPR TC-12 Dataset) Using the PRF-SF Method

Table 23: Comparison between the Diversity of Text-only and PRF-SF Results on the Belga Set Using Cluster Recall

| Method | CR@5 | CR@10 | CR@20 | CR@30 | CR@100 |
|---|---|---|---|---|---|
| Text-only | 0.4974 | 0.6033 | 0.6645 | 0.7078 | 0.8784 |
| PRF-SF | 0.5191 | 0.6095 | 0.7205 | 0.7918 | 0.9119 |

## 5.7 Examples of Query Expansions

This section illustrates concrete examples of the query expansion, analyzing the conditions for a successful query expansion and those that are disadvantageous.

### 5.7.1 Part Meronym-Filtered Expansions

We first illustrate with query expansion filtered using the "Part Meronym" relation from the WordNet database. The original query "Straight Road in the USA" is from the 2008 ImageCLEFPhoto benchmark which was shown in Figure 10. The final query after the expansion and the weight assignment is shown in Figure 35:

```
usa^9  straight road highway traffic Unite State America highway^4  colorado^4
america^4 cyclist^0.9 ride^2 grey^0.9 road^4 flat^2 landscap^2 green^0.9
meadow^0.9 dark^0.9 red^0.9 car^0.9 power^0.9 pole^2 right^2 blue^0.9 sky^0.9
white^0.9 cloud^0.9 background^2 highway^4 160^0.9 kansas^4 america^4  cyclist^0.9
red^0.9 blue^0.9 white^0.9 jersey^0.9 black^0.9 cycl^0.9 short^2 blue^0.9 helmet^0.9
ride^2 red^0.9 black^0.9 race^0.9 bike^0.9 grey^0.9 road^4 flat^2 landsap^0.9 green^0.9
meadow^0.9 dark^0.9 red^0.9 car^0.9 sponsor^0.9 sticker^0.9 spare^0.9 frame^2 spare^0.9
wheel^0.9 yellow^0.9 signage unit^0.9 roof^0.9 rack^2 power^0.9 pole^2 left^2 white^0.9
car^0.9 greyish blu^0.9 sky^0.9 background^2
```

Figure 35: Query "Straight Road in the USA" After Expansion

As can be seen from the expansion, the terms from the top visual results used for expansion of the query which were found in the Part-Meronym hierarchy of the term USA were given more weight (Colorado, America, Kansas). The original query term "USA" was found in other top visual results and so it was given more weight (9). Terms that were found in the related lists built from WordNet relation were given twice the weight, and the other expansion terms were given a weight less than 1 (0.9). This promotes recall without significantly impacting the precision. For these expansion terms for which no relation to the original query

can be found, we experimented with weights ranging from 0.2 to 0.9, and did not find major differences in the results.

### 5.7.2 Successful Expansions

Figures 36, 37, 38 and 39 demonstrate the queries with expansions that resulted in the most significant improvements in the Mean Average Precision (MAP) over the text-only runs using the ImageCLEFPhoto 2009 queries on the Belga dataset in order of the magnitude of improvement. The first of these figures, Figure 36 shows the expanded query with the most staggering improvement over the text-only method, achieving a MAP of 0.5833 compared to 0.0181 for text retrieval. The query expansion shown in the following two figures (Figure 37 and Figure 38) also resulted in very significant improvements (0.3455 vs 0.8486 and 0.4614 vs. 0.9078 respectively). The last query expansion illustrated in Figure 39, while not as successful measured in absolute terms, produced a MAP of 0.283 compared to a MAP of 0.0593 for the text-only retrieval method. These examples prove that using the visual input with query expansion has the potential of producing meaningful results to queries that a text-only method is incapable of handling in a meaningful manner.

```
epa^0.5 ellen^5 degeneres^5 host^0.5 th^0.5 annual^0.5 academi^0.5 award^0.5
kodak^0.5 theatr^0.5 hollywood^0.5 ca^0.5 epa michael^0.5 yada^0.5
a m p a s^0.5 editorial^0.5 use^0.5 time^0.5 use^0.5 sales^0.5 archives^0.5
```

Figure 36: Query Expansion for the Query “ellen degeneres”

```
brussels^0.5 belgium^0.5 illustrat^0.5 show^0.5 new^0.5 logo fortis^5 bank^0.5
insur^0.5 group^0.5
```

Figure 37: Query Expansion for the Query “Fortis”

### 5.7.3 Noisy Expansions

While Section 5.7.2 demonstrated the improvements that query expansion can achieve over text-only retrieval, there are cases where the query expansion introduced to much noise in the query that was not handled properly by the semantic filters. This section analyzes the most significant of these cases.

```
fernand Fernand Koekelberg Fernand peopl shown brussels^0.5 belgium^0.5 new^0.5
polic^0.5 superintend^0.5 fernand^0.5 koekelberg^5 pictur^0.5 took^0.5 oath^0.5
superintend^0.5 brussels^0.5 belgium^0.5 new^0.5 polic^0.5 superintend^0.5
fernand^0.5 koekelberg^5 pictur^0.5 took^0.5 oath^0.5 superintend^0.5
```

Figure 38: Query Expansion for the Query “koekelberg”

```
bru^0.5 verviers^0.5 belgium^0.5 olivia^0.5 borlee^5 member belgian^0.5 athlet team
receiv^0.5 sportif^0.5 merit^0.5 price^0.5 french^0.5 communiti^0.5 hope^0.5
vervier^0.5 michel^0.5 krakowski^0.5
```

Figure 39: Query Expansion for the Query “olivia borlee”

Only one query suffered from a major deterioration in the MAP obtained due to query expansion. The topic of that query is *monaco*, with the cluster *albert monaco*, *stephanie monaco*, and *caroline monaco*. Figure 40 shows the resulting expansion from the different example images. The Map achieved by the original text query is 0.5478 while that obtained by the expanded query is 0.2905. One factor that can explain this important deterioration is the length of the documents from which the query was expanded. Longer documents resulted in longer queries, since we did not take into consideration limiting the expansion to specific region of the document.

```
/belga31/06178527.jpg:epa^0.5 princ^0.5 albert ii^0.5 monaco^5 fight^0.5 aid^0.5
monaco^5 presid^0.5 honor^0.5 attend^0.5 chariti^0.5 gala^0.5 grimaldi^0.5
forum^0.5 monaco^5 event^0.5 join^0.5 music^0.5 sun^0.5 king^0.5 relat^0.5
life^0.5 french^0.5 king^0.5 loui^0.5 xiv^0.5 philharmon^0.5 orchestra^0.5
monaco^5 number^0.5 french^0.5 internat^0.5 renown^0.5 artist^0.5 play^0.5
sall^0.5 des^0.5 princ^0.5 monaco^5 profit^0.5 go^0.5 fight^0.5 aid^0.5
monaco^5 epa asm^0.5 corbis^0.5 15  monaco  princ albert Princ Albert Monaco
Princ Albert peopl shown epa^0.5 princ^0.5  albert ii^0.5 monaco^5 fight^0.5
aid^0.5 monaco^5 presid^0.5 honor^0.5 attend^0.5 chariti^0.5 gala^0.5
```

Figure 40: Noisy Expansion of the Query “prince albert of monaco”

# Chapter 6

# Conclusion and Future Work

This chapter concludes the dissertation by pinpointing the advantages of the proposed method in comparison to text-only retrieval methods as well as different multi-modal retrieval methods. Section 6.1 starts the chapter, recapitulating the salient points of our approach, Finally, Section 6.3 suggests possible directions of research that can benefit from our approach, and build on it.

## 6.1 The Proposed Method in a Nutshell

Research in the field of photographic image retrieval has been recently achieving more success with methods that incorporate both textual and visual data available. Major challenges in this area include handling the very high resource-bound visual (content-based) retrieval, and bridging the semantic gap between the visual features and the meaning of the content. This dissertation proposes several methods to overcome these obstacles. The method we propose relies mainly on Pseudo-Relevance Feedback and Semantic Filtering (PRF-SF), starting from visual retrieval. For the visual retrieval component, a block-based method is proposed which is relatively less resource intensive than sophisticated visual similarity techniques, and which achieves better early precision. The resulting documents are then exploited for semantically-filtered query expansion, which takes into consideration the potential of semantic relatedness between the documents and the query to add terms to the original query, as well as assign them appropriate weights. The method is validated

by applying them to two ImageCLEF benchmark diverse datasets: the IAPR TC-12 collection consisting of tourist photographs, and the Belga news agency collection. Results with these benchmarks demonstrate that the method combining the visual retrieval component with the semantic constraints on the expansion is indeed effective in improving the original query through pseudo-relevance feedback, and are superior to other methods used in the benchmarks. Other questions investigated in this thesis include the effect of the size and nature of the data collection on the results (Section 4.1.7), the different visual features used (Section 4.2), and employing different retrieval models for the text retrieval component (Section 4.1.4), in addition to different semantic relations for filters on the query expansion, and additional fusion of the results.

## 6.2 Research Contributions

Five angles of research were investigated within the scope of Image Retrieval in this dissertation:

- Text-based Retrieval
- Content-based Retrieval
- Fusion Methods for Text-based and Content-based Retrieval
- Query Expansion Employing Semantic Filters
- Pseudo-Relevance Feedback

Following is a summary of the main findings of the research conducted.

- For the first angle of research, no significant differences were found between the text retrieval models experimented with within the context of image retrieval task. This is due to the scarcity of the text. We also demonstrated that the availability of more text significantly improves the precision of results.
- For the second angle, we developed a block-based method which is not resource-intensive for visual retrieval. This method outperforms MPEG-7 descriptors especially at early recall, and is therefore suitable for Pseudo-Relevance Feedback.

- For the third angle, we investigated several ways for fusion starting from both text and image results. Mean Average Precision (MAP), recall and cluster recall can all be improved by using using inter-media fusion methods such as PRF-SF. We have also shown that incorporating both text and visual features promotes diversity in image retrieval results

- For the fourth angle, we developed a robust method for inter-media query expansion using semantic filtering of text that was tested on different datasets and demonstrated the feasibility of query expansion that are not noisy.

- For the final angle, the developed PRF-SF method incorporates pseudo-relevance feedback starting from visual retrieval results.

## 6.3 Further Research Directions

To conclude the dissertation, we would like to suggest directions of research that can benefit from the research carried out within its scope:

- Context-aware image retrieval: Evidently, since our PRF-SF method is capable of tackling visual examples, it lends itself to the sphere of relevance feedback within a context-aware image retrieval framework. Given the rapid advances in hardware, improvements can be made to the simple visual retrieval method we used while minimizing the effects on the execution speed.

- Long documents: An important factor to investigate in real-life scenarios is the region of the document to use for query expansion in case of long documents. This would avoid problems such as the one described in Section 5.7.3. An example can be found in [Coelho *et al.*, 2004], who experimented with different sizes of text passages as well as combinations of text passages and HTML metatags on a collection of 54,000 images from the Brazilian web.

- Post-retrieval fusion and filtering: Finally, while the pre-querying semantic filtering applied in our approach can be useful, combining it with a more sophisticated post-retrieval filtering than we applied

in order to remove noise and confirm the relevance of the results, could potentially further improve the result.

- Evaluation with Other Datasets: Although the PRF-SF method has been tested and validated using two collections of different sizes and sources, further evaluation on even larger datasets can shed more light on the applicability of the method.

# Appendices

# Appendix A

# List of Stop words Used

a about above across after afterwards again against all almost alone along already also although always am among amongst amoungst amount an and another any anyhow anyone anything anyway anywhere are around as at back be became because become becomes becoming been before beforehand behind being below beside besides between beyond bill both bottom but by call can cannot cant co computer con could couldnt cry de describe detail do done down due during each eg eight either eleven else elsewhere empty enough etc even ever every everyone everything everywhere except few fifteen fify fill find fire first five for former formerly forty found four from front full further get give go had has hasnt have he hence her here hereafter hereby herein hereupon hers herself him himself his how however hundred i ie if image images in inc indeed interest into is it its itself keep last latter latterly least less ltd made many may me meanwhile might mill mine more moreover most mostly move much must my myself name namely neither never nevertheless next nine no nobody none noone nor not nothing now nowhere of off often on once one only onto or other others otherwise our ours ourselves out over own part per perhaps photograph photographs please put rather re relevant Relevant same see seem seemed seeming seems serious several she should show side since sincere six sixty so some somehow someone something sometime sometimes somewhere still such system take ten than that the their them themselves then thence there thereafter thereby therefore therein thereupon these they thick thin third this those though three through throughout thru thus to together too top toward towards twelve

twenty two un under until up upon us very via was we well were what whatever when whence whenever where whereafter whereas whereby wherein whereupon wherever whether which while whither who whoever whole whom whose why will with within without would yet you your yours yourself yourselves monday tuesday wednesday thursday friday saturday sunday january february march april may june july august september october november december contain match foreground cluster picture pictur categori category belga includ include photo

# Bibliography

[Agosti *et al.*, 2010] M. Agosti, N. Ferro, C. Peters, M. de Rijke, and A.F. Smeaton, editors. *Multilingual and Multimodal Information Access Evaluation, International Conference of the Cross-Language Evaluation Forum, CLEF 2010*, volume 6360 of *Lecture Notes in Computer Science*, Padua, Italy, September 20-23 2010. Springer.

[Agrawal *et al.*, 2009] Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, and Samuel Ieong. Diversifying search results. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 5–14, New York, NY, USA, 2009. ACM.

[Armitage and Enser, 1997] L. Armitage and P. Enser. Analysis of user need in image archives. *Journal of Information Science*, pages 287–299, 1997.

[Arni *et al.*, 2008 printed in 2009] Thomas Arni, Paul Clough, Mark Sanderson, and Michael Grubinger. Overview of the ImageCLEFphoto 2008 photographic retrieval task. In Carol Peters, Danilo Giampiccol, Nicola Ferro, Vivien Petras, Julio Gonzalo, Anselmo Peñas, Thomas Deselaers, Thomas Mandl, Gareth Jones, and Mikko Kurimo, editors, *Evaluating Systems for Multilingual and Multimodal Information Access – 9th Workshop of the Cross-Language Evaluation Forum*, Lecture Notes in Computer Science, Aarhus, Denmark, 2008 (printed in 2009).

[Aslam *et al.*, 2005] Javed A. Aslam, Emine Yilmaz, and Virgiliu Pavlu. A geometric interpretation of r-precision and its correlation with average precision. In *Proceedings of the 28th Annual International ACM*

*SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, pages 573–574, New York, NY, USA, 2005. ACM.

[Bartolini, 2005] Ilaria Bartolini. Context-based image similarity queries. In *Proceedings of the Third International Workshop on Adaptive Multimedia Retrieval (AMR 2005)*, Glasgow, UK, 2005.

[Besançon and Millet, 2005] Romaric Besançon and Christophe Millet. Merging Results from Different Media: Lic2m experiments at ImageCLEF 2005. In *Working Notes of the CLEF Workshop*, Vienna, Austria, 21-23 September 2005.

[Bird *et al.*, 2009] Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly, Beijing, 2009.

[Brill, 1992] Eric Brill. A simple rule-based part of speech tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing*, ANLC '92, pages 152–155, Stroudsburg, PA, USA, 1992. Association for Computational Linguistics.

[Buckley and Voorhees, 2000] Chris Buckley and Ellen M. Voorhees. Evaluating evaluation measure stability. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '00, pages 33–40, New York, NY, USA, 2000. ACM.

[Budzik and Hammond, 2000] J. Budzik and K. Hammond. User interactions with everyday applications as context fot just-in-time information access. In *IUI'2000, Proceedings of the 2000 International Conference on Intelligent User Interfaces*, pages 44–51. ACM, 2000.

[Cai *et al.*, 2004a] Deng Cai, Xiaofei He, Zhiwei Li, Wei-Ying Ma, and Ji-Rong Wen. Hierarchical clustering of www image search results using visual, textual and link information. In *MULTIMEDIA '04: Proceedings of the 12th Annual ACM International Conference on Multimedia*, pages 952–959, New York, NY, USA, 2004. ACM Press.

[Cai *et al.*, 2004b] Deng Cai, Xiaofei He, Wei-Ying Ma, Ji-Rong Wen, and HongJiang Zhang. Organizing WWW images based on the analysis of page layout and web link structure. In *Proceedings of the 2004*

*IEEE International Conference on Multimedia and Expo (ICME)*, pages 113–116, Taipei,Taiwan, 27-30 June 2004 2004. IEEE.

[Cascia *et al.*, 1998] Marco La Cascia, Saratendu Sethi, and Stan Sclaroff. Combining textual and visual cues for content-based image retrieval on the world wide web. In *IEEE Workshop on Content-based Access of Image and Video Libraries*, pages 24–28, 1998.

[Chang *et al.*, 2005] Yih-Cheng Chang, Wen-Cheng Lin, and Hsin-Hsi Chen. Combining text and image queries at ImageClef 2005. In *Working notes of the CLEF workshop*, Vienna, Austria, 2005.

[Chen *et al.*, 2003] Yixin Chen, James Z. Wang, and Robert Krovetz. Content-based image retrieval by clustering. In *MIR '03: Proceedings of the 5th ACM SIGMM international workshop on Multimedia information retrieval*, pages 193–200, New York, NY, USA, 2003. ACM Press.

[Clarke *et al.*, 2011a] Charles L. A. Clarke, Nick Craswell, Ian Soboroff, and Ellen M. Voorhees. Overview of the trec 2011 web track. In Ellen M. Voorhees and Lori P. Buckland, editors, *TREC*. National Institute of Standards and Technology (NIST), 2011.

[Clarke *et al.*, 2011b] Charles L.A. Clarke, Nick Craswell, Ian Soboroff, and Azin Ashkan. A comparative analysis of cascade measures for novelty and diversity. In *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM '11, pages 75–84, New York, NY, USA, 2011. ACM.

[Clinchant *et al.*, 2010] Stéphane Clinchant, Gabriela Csurka, Julien Ah-Pine, Guillaume Jacquet, Florent Perronnin, Jorge Sánchez, and Keyvan Minoukadeh. Xrce's participation in Wikipedia retrieval, medical image modality classification and ad-hoc retrieval tasks of Imageclef 2010. In Agosti et al. [2010].

[Clough *et al.*, 2005a] P. Clough, H. Joho, and M. Sanderson. Automatically Organising Images using Concept Hierarchies. Workshop held at the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Workshop: Multimedia Information Retrieval In Salvador, Brazil, August 15-19 2005.

[Clough *et al.*, 2005b] Paul Clough, Henning Mller, and Mark Sanderson. The clef 2004 cross language image retrieval track. In *Multilingual Information Access for Text, Speech and Images: Results of the Fifth CLEF Evaluation Campaign, Lecture Notes In Computer Science*, page 2005. Springer, 2005.

[Coelho *et al.*, 2004] Tatiana Almeida Souza Coelho, Pvel Pereira Calado, Lamarque Vieira Souza, Berthier Ribeiro-Neto, and Richard Muntz. Image retrieval using multiple evidence ranking. *IEEE Transactions on Knowledge and Data Engineering*, 16(4):408–417, 2004.

[Cumbreras *et al.*, 2009] Miguel Ángel García Cumbreras, Manuel Carlos Díaz-Galiano, Maria Teresa Martín-Valdivia, Arturo Montejo Ráez, and Luis Alfonso Ureña López. University of Jaén at ImageCLEF 2009: Medical and photo tasks. In Peters et al. [2010], pages 348–353.

[Datta *et al.*, 2008] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Comput. Surv.*, 40(2):1–60, May 2008.

[de Marneffe *et al.*, 2006] Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. Generating Typed Dependency Parses from Phrase Structure Parses. In *Proceedings of the IEEE / ACL 2006 Workshop on Spoken Language Technology*. The Stanford Natural Language Processing Group, 2006.

[Demerdash *et al.*, 2009] Osama El Demerdash, Sabine Bergler, and Leila Kosseim. Image query expansion using semantic selectional restrictions. In Peters et al. [2010], pages 150–156.

[Deselaers and Hanbury, 2009] Thomas Deselaers and Allan Hanbury. The visual concept detection task in ImageCLEF 2008. In *CLEF Workshop 2008 / Evaluating Systems for Multilingual and Multimodal Information Access*, volume 5706 of *LNCS*, Aarhus, Denmark, 2009. Springer.

[Deselaers *et al.*, 2004] T. Deselaers, D. Keysers, and H. Ney. FIRE — Flexible Image Retrieval Engine: ImageCLEF 2004 evaluation. In *CLEF Workshop (2004)*, 2004.

[Douze *et al.*, 2009] Matthijs Douze, Hervé Jégou, Harsimrat Sandhawalia, Laurent Amsaleg, and Cordelia Schmid. Evaluation of GIST descriptors for web-scale image search. In *Proceeding of the ACM International Conference on Image and Video Retrieval*, CIVR '09, pages 19:1–19:8, New York, NY, USA, 2009. ACM.

[Douze *et al.*, 2011] Matthijs Douze, Arnau Ramisa, and Cordelia Schmid. Combining attributes and Fisher vectors for efficient image retrieval. In *IEEE Conference on Computer Vision & Pattern Recognition, CVPR 2011*. IEEE, June 2011.

[El Demerdash *et al.*, 2008] Osama El Demerdash, Leila Kosseim, and Sabine Bergler. Text-based clustering of the ImageCLEFPhoto collection for augmenting the retrieved results. In *Advances in Multilingual and Multimodal Information Retrieval: 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Revised Selected Papers. LNCS*, Lecture Notes in Computer Science, Volume 5152, pages 562–568, Budapest, Hungary, September 19-21 2008.

[El Demerdash *et al.*, 2009a] Osama El Demerdash, Sabine Bergler, and Leila Kosseim. CLaC at imageclef 2009. In *CLEF working notes 2009*, Corfu, Greece, 2009.

[El Demerdash *et al.*, 2009b] Osama El Demerdash, Leila Kosseim, and Sabine Bergler. Image retrieval by inter-media fusion and pseudo-relevance feedback. In *Proceedings of the 9th Cross-Language Evaluation Forum Conference on Evaluating Systems for Multilingual and Multimodal Information Access, CLEF'08*, volume 5706 of *LNCS*, pages 605–611, Berlin, Heidelberg, 2009. Springer-Verlag.

[El Demerdash *et al.*, 2009c] Osama El Demerdash, Leila Kosseim, and Sabine Bergler. Semantic inter-media image retrieval in photographic collections. In *Multimedia Signal Processing(MMSP)*, pages 1–5. IEEE, 2009.

[Fakeri-Tabrizi *et al.*, 2010] Ali Fakeri-Tabrizi, Sabrina Tollari, Nicolas Usunier, and Patrick Gallinari. UPMC/LIP6 at ImageCLEFAnnotation 2010. In *CLEF (Notebook Papers/LABs/Workshops)*, 2010.

[Fauqueur and Boujemaa, 2004] Julien Fauqueur and Nozha Boujemaa. Mental image search by boolean composition of region categories. *Multimedia Tools and Applications*, 31:95–117, 2004.

[Fellbaum, 1998] Christiane Fellbaum, editor. *WordNet An Electronic Lexical Database*. The MIT Press, Cambridge, MA ; London, May 1998.

[Flickner *et al.*, 1997] Myron Flickner, Harpreet Sawhney, and Wayne Nublack. *Intelligent Multimedia Information Retrieval*, chapter Query by Image and Video Content: The QBIC System. California:AAAI Press/ The MIT Press, 1997.

[Gao *et al.*, 2005] Bin Gao, Tie-Yan Liu, Tao Qin, Xin Zheng, Qian-Sheng Cheng, and Wei-Ying Ma. Web image clustering by consistent utilization of visual features and surrounding texts. In *MULTIMEDIA '05: Proceedings of the 13th Annual ACM International Conference on Multimedia*, pages 112–121, New York, NY, USA, 2005. ACM Press.

[Garber and Grunes, 1992] Sharon R. Garber and Mitch B. Grunes. The art of search: A study of art directors. In *CHI '92: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 157–163, New York, NY, USA, 1992. ACM Press.

[Gevers and Smeulders, 2000] T. Gevers and A. W. M. Smeulders. Pictoseek: Combining color and shape invariant features for image retrieval. *IEEE Transactions on Image Processing*, 9(1):102–119, 2000.

[Gonzalo *et al.*, 2005] J. Gonzalo, P. Clough, and A. Vallin. Overview of the CLEF 2005 interactive track. In *Working notes of the CLEF workshop*, Vienna, Austria, 21-23 September 2005.

[Goodrum and Spink, 1999] A. Goodrum and A. Spink. Visual information seeking: A study of image queries on the world wide web. In *Proceedings of the 1999 Annual meeting of the American Society for Information Science*, Washington, DC, USA, 1999.

[Goodrum, 2000] A. Goodrum. Image information retrieval: An overview of current research. *Informing Science*, 3(2):63–66, 2000.

[Google, 2006] Google. Google search engine, 2006. [Online; accessed 14-April-2006].

[Goren-Bar *et al.*, 2001] D. Goren-Bar, T.Kuflik, and T. Lavie. What do users prefer? a personalized intelligent user interface for searching information - an empirical study. In James C. Lester, editor, *IUI 2001 -*

*Proceedings of the 2001 International Conference on Intelligent User Interfaces*, pages 65–68, Santa Fe, New Mexico, January 2001. ACM Press.

[Gruber, 1993] Thomas R. Gruber. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2):199–220, 1993.

[Grubinger *et al.*, 2005] M. Grubinger, C. Leung, and P. Clough. The IAPR benchmark for assessing image retrieval performance in cross language evaluation tasks. In *Proceedings of the first MUSCLE / ImageCLEF workshop on image and video retrieval evaluation*, Vienna, Austria, 20th September 2005.

[Grubinger *et al.*, 2007] Michael Grubinger, Paul Clough, Allan Hanbury, and Henning Müller. Overview of the ImageCLEFphoto 2007 Photographic Retrieval Task. In *Working Notes of the 2007 CLEF Workshop*, Budapest, Hungary, September 2007.

[Han and Huang, 2005] Jun-Hua Han and De-Shuang Huang. A novel BP-Based Image Retrieval System. In *International Symposium on Circuits and Systems (ISCAS 2005)*, pages 1557–1560, Kobe, Japan, 23-26 May 2005. IEEE.

[Harper and Hendry, 1997] David Harper and David Hendry. Evaluation light. In M.D. Dunlop, editor, *Proceedings of the Second Mira Workshop*, University of Glasgow, Computing Science Research Report TR-1997-2, Monselice, Italy, 1997.

[Hatcher and Gospodnetic, 2004] Erik Hatcher and Otis Gospodnetic. *Lucene in Action (In Action series)*. Manning Publications Co., Greenwich, CT, USA, 2004.

[Heesch and Rüger, 2008] Daniel Heesch and Stefan Rüger. Two step relevance feedback for semantic disambiguation in image retrieval. In *VISUAL'08: Proceedings of the 10th international conference on Visual Information Systems*, pages 204–215, Berlin, Heidelberg, 2008. Springer-Verlag.

[Hoi *et al.*, 2005] Steven C.H. Hoi, Jianke Zhu, and Michael R. Lyu. CHUK experiments with ImageCLEF 2005. In *working notes of the CLEF workshop*, Vienna, Austria, 21-23 September 2005.

[Huang *et al.*, 2001] Lieming Huang, Thiel Ulrich, and Matthias Hemmje. Adaptively constructing the query interface for meta-search engines. In *Intelligent User Interfaces, IUI'01*, pages 97–100. ACM, 2001.

[Huijsmans and Sebe, 2005] Dionysius P. Huijsmans and Nicu Sebe. How to complete performance graphs in content-based image retrieval: Add generality and normalize scope. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27:245–251, 2005.

[Ingwersen and Järvelin, 2005] Peter Ingwersen and Kalervo Järvelin. Information retrieval in context: Irix. *SIGIR Forum*, 39(2):31–39, 2005.

[Inkpen *et al.*, 2009] Diana Inkpen, Marc Stogaitis, François DeGuire, and Muath Alzghool. Clustering for photo retrieval at image clef 2008. In *Proceedings of the 9th Cross-language evaluation forum conference on Evaluating systems for multilingual and multimodal information access*, CLEF'08, pages 685–690, Berlin, Heidelberg, 2009. Springer-Verlag.

[Jain and Dubes, 1988] Anil K. Jain and Richard C. Dubes. *Algorithms for clustering data*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1988.

[Jain *et al.*, 1999] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Computing Surveys*, 31(3):264–323, 1999.

[Jones, 1972] Karen Spärck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:11–21, 1972.

[Juffinger *et al.*, 2009] Andreas Juffinger, Roman Kern, and Michael Granitzer. Crosslanguage retrieval based on Wikipedia statistics. In *Proceedings of the 9th Cross-language evaluation forum conference on Evaluating systems for multilingual and multimodal information access*, CLEF'08, pages 155–162, Berlin, Heidelberg, 2009. Springer-Verlag.

[Kullback and Leibler, 1951] S. Kullback and A. Leibler. On information and sufficiency. *IEEE Transactions on Information Theory*, 22:79–86, 1951.

[Leake and Scherle, 2001] David B. Leake and Ryan Scherle. Towards context-based search engine selection. In *IUI'01, Proceedings of the 2001 International Conference on Intelligent User Interfaces*, pages 109–112. ACM, 2001.

[Leouski and Croft, 1996] A. Leouski and W. Croft. An evaluation of techniques for clustering search results. Technical Report IR-76, Department of Computer Science, University of Massachusetts, Amherst, 1996.

[Lew *et al.*, 2006] Michael S. Lew, Nicu Sebe, Chabane Djeraba, and Ramesh Jain. Content-based multimedia information retrieval: State of the art and challenges. *ACM Trans. Multimedia Comput. Commun. Appl.*, 2:1–19, February 2006.

[Li *et al.*, 2005] Zhiwei Li, Gu Xu, Mingjing Li, Wei-Ying Ma, and Hong-Jiang Zhang. Grouping WWW image search results by novel inhomogeneous clustering method. In Yi-Ping Phoebe Chen, editor, *11th International Conference on Multi Media Modeling (MMM 2005)*, pages 255–261. IEEE Computer Society, 2005.

[Liu *et al.*, 2007] Ying Liu, Dengsheng Zhang, Guojun Lu, and Wei-Ying Ma. A survey of content-based image retrieval with high-level semantics. *Pattern Recogn.*, 40:262–282, January 2007.

[Lux and Granitzer, 2005] Mathias Lux and Michael Granitzer. Retrieval of MPEG-7 Based Semantic Descriptions. In *BTW-Workshop WebDB Meets IR at the GI-Fachtagung für Datenbanksysteme in Business*, Technologie und Web, University Karlsruhe, 2005.

[Magesh and Thangaraj, 2011] N. Magesh and Dr. P. Thangaraj. Article: Semantic image retrieval based on ontology and sparql query. *International Journal of Computer Applications*, ICACT(1):12–16, August 2011. Published by Foundation of Computer Science, New York, USA.

[Maillot *et al.*, 2006] Nicolas Maillot, Jean-Pierre Chevallet, and Joo-Hwee Lim. Inter-media pseudo-relevance feedback application to ImageCLEF 2006 photo retrieval. In Carol Peters, Paul Clough, Fredric C. Gey, Jussi Karlgren, Bernardo Magnini, Douglas W. Oard, Maarten de Rijke, and Maximilian Stempfhuber, editors, *CLEF*, volume 4730 of *Lecture Notes in Computer Science*, pages 735–738. Springer, 2006.

[Mandal *et al.*, 1996] M.K. Mandal, T. Aboulnasr, and S. Panchanathan. Image indexing using moments and wavelets. *IEEE Transactions on Consumer Electronics*, 42:557–565, 1996.

[Manning *et al.*, 2008] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008. Online 17/08/2007. http://www-csli.stanford.edu/~schuetze/information-retrieval-book.html.

[Markkula and Sormunen, 2000] Marjo Markkula and Eero Sormunen. End-user searching challenges indexing practices inthe digital newspaper photo archive. *Information Retrieval*, 1(4):259–285, 2000.

[Martínez-Fernández *et al.*, 2005] J.L. Martínez-Fernández, J. Villena, Ana García-Serrano, S. Gonzalez-Tortosa, F.Carbone, and M. Castagnone. Exploiting semantic features for image retrieval at CLEF 2005. In *working notes of the CLEF workshop*, Vienna, Austria, 21-23 September 2005.

[Martínez-Fernández *et al.*, 2006] J.L. Martínez-Fernández, Julio Román, Ana García-Serrano, and José González-Cristóbal. Combining Textual and Visual Features for Image Retrieval. *Accessing Multilingual Information Repositories*, pages 680–691, 2006.

[Martínez, 2004] José Martínez. MPEG-7 Overview (version 10). Technical Report N6828, ISO/IEC JTC1/SC29/WG11 (MPEG), October 2004. online August 17, 2007 http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm.

[MSN, 2006] MSN. MSN search engine, 2006. [Online; accessed 14-April-2006].

[Müller *et al.*, 2002] Henning Müller, Stéphane Marchand-Maillet, and Thierry Pun. The truth about corel - evaluation in image retrieval. In *Proceedings of the International Conference on Image and Video Retrieval*, CIVR '02, pages 38–49, London, UK, 2002. Springer-Verlag.

[Müller *et al.*, 2005a] H. Müller, P. Clough, A. Geissbuhler, and W. Hersh. ImageCLEF 2004-2005: Results, experiences and new ideas for image retrieval evaluation. In *Proceedings of the Fourth International Workshop on Content-Based Multimedia Indexing (CBMI2005), Riga, Latvia*, 2005.

[Müller *et al.*, 2005b] H. Müller, P. Clough, W. Hersh, T. Deselaers, T. Lehmann, and M. Grubinger. Overview of the 2005 cross-language image retrieval track (ImageCLEF). In *working notes of the CLEF workshop*, Vienna, Austria, 21-23 September 2005.

[Müller *et al.*, 2006] Henning Müller, Paul Clough, William Hersh, Thomas Deselaers, Thomas Lehmann, and Antoine Geissbuhler. Using heterogeneous annotation and visual information for the benchmarking of image retrieval systems. In *SPIE Conference Photonics West, Electronic Imaging, special session on benchmarking image retrieval systems*, San Diego, February 2006.

[Oliva and Torralba, 2001] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42:145–175, 2001.

[Osinski and Weiss, 2005] Stanislaw Osinski and Dawid Weiss. Carrot$^2$: Design of a flexible and efficient web information retrieval framework. In Piotr S. Szczepaniak, Janusz Kacprzyk, and Adam Niewiadomski, editors, *AWIC*, volume 3528 of *Lecture Notes in Computer Science*, pages 439–444. Springer, 2005.

[Ounis *et al.*, 2006] I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, and C. Lioma. Terrier: A High Performance and Scalable Information Retrieval Platform. In *Proceedings of ACM SIGIR'06 Workshop on Open Source Information Retrieval (OSIR 2006)*, 2006.

[Paramita *et al.*, 2009] M. Paramita, M. Sanderson, and P. Clough. Diversity in Photo Retrieval: Overview of the ImageCLEFPhoto Task 2009. In *CLEF working notes 2009, Corfu, Greece*, 2009.

[Peters *et al.*, 2010] Carol Peters, Barbara Caputo, Julio Gonzalo, Gareth J. F. Jones, Jayashree Kalpathy-Cramer, Henning Müller, and Theodora Tsikrika, editors. *Multilingual Information Access Evaluation II. Multimedia Experiments - 10th Workshop of the Cross-Language Evaluation Forum, CLEF 2009, September 30 - October 2, 2009, Revised Selected Papers*, volume 6242 of *LNCS*, Corfu, Greece, 2010. Springer.

[Petrelli and Clough, 2005] D. Petrelli and P. Clough. Concept hierarchy across languages in text-based image retrieval: A user evaluation. In *Working notes for the CLEF 2005 workshop*, Vienna, Austria, 2005.

[Pharo and Järvelin, 2006] Nils Pharo and Kalervo Järvelin. Irrational searchers and IR-rational researchers. *Journal of the American Society for Information Science and Technology*, 57(2):222–232, 2006.

[Pharo, 2004] Nils Pharo. A new model of information behaviour based on the search situation transition schema. *Information Research*, 10(1), 2004.

[Popescu *et al.*, 2007] Adrian Popescu, Christophe Millet, and Pierre-Alain Moëllic. Ontology driven content based image retrieval. In *Proceedings of the 6th ACM International Conference on Image and Video Retrieval*, CIVR '07, pages 387–394, New York, NY, USA, 2007. ACM.

[Porter, 2001] Martin F. Porter. Snowball: A language for stemming algorithms. Published online, October 2001. Accessed 16.04.2009, 18.00h.

[Qiu, 2004] Guoping Qiu. Image and feature co-clustering. In *ICPR '04: Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04) Volume 4*, pages 991–994, Washington, DC, USA, 2004. IEEE Computer Society.

[Quack *et al.*, 2004] Till Quack, Ullrich Mönich, Lars Thiele, and B. S. Manjunath. Cortina: A system for large-scale, content-based web image retrieval. In *Proceedings of the 12th Annual ACM International Conference on Multimedia*, MULTIMEDIA '04, pages 508–511, New York, NY, USA, 2004. ACM.

[Ratinov and Roth, 2009] Lev Ratinov and Dan Roth. Design challenges and misconceptions in named entity recognition. In *CoNLL '09: Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pages 147–155, Morristown, NJ, USA, 2009. Association for Computational Linguistics.

[Reid, 1999] N.H. Reid. *The photographic collections in St Andrews University Library*, volume 5 of *Scottish Archives*, pages 83–90. 1999.

[Robertson and Sparck Jones, 1976] S. E. Robertson and K. Sparck Jones. Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27(3):129–146, 1976.

[Robertson and Walker, 1994] S. E. Robertson and S. Walker. Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th annual international ACM*

*SIGIR conference on research and development in information retrieval*, SIGIR '94, pages 232–241, New York, NY, USA, 1994. Springer-Verlag New York, Inc.

[Robertson and Zaragoza, 2009] Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3:333–389, April 2009.

[Robertson *et al.*, 1996] S.E. Robertson, S. Walker, S. Jones, M.M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. In *Overview of the Third Text REtrieval Conference (TREC3)*, pages 109–126. Gaithersburg, MD: NIST, 1996.

[Rodden *et al.*, 2001] Kerry Rodden, Wojciech Basalaj, David Sinclair, and Kenneth Wood. Does organisation by similarity assist image browsing? In *CHI '01: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 190–197, New York, NY, USA, 2001. ACM Press.

[Rose and Levinson, 2004] Daniel E. Rose and Danny Levinson. Understanding user goals in web search. In *WWW '04: Proceedings of the 13th international conference on World Wide Web*, pages 13–19, New York, NY, USA, 2004. ACM Press.

[Rui and Huang, 1999] Yong Rui and Thomas S. Huang. Image retrieval: Current techniques, promising directions and open issues. *Journal of Visual Communication and Image Representation*, 10:39–62, 1999.

[Rui *et al.*, 1997] Yong Rui, Thomas S. Huang, and Sharad Mehrotra. Content-based image retrieval with relevance feedback in MARS. In *International Conference on Image Processing(ICIP)(2)*, pages 815–818, 1997.

[Salton *et al.*, 1975] Gerard Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620, 1975.

[Sivic and Zisserman, 2003] Josef Sivic and Andrew Zisserman. Video Google: A text retrieval approach to object matching in videos. In *Proceedings of the Ninth IEEE International Conference on Computer Vision - Volume 2*, ICCV '03, pages 1470–1477, Washington, DC, USA, 2003. IEEE Computer Society.

[Smeaton, 2001] A.F. Smeaton. The TREC 2001 video track report. In E.M. Voorhees and D.K. Harman, editors, *The Tenth Text Retrieval Conference, TREC 2001*, NIST Special Publication 500-250, pages 52–60. NIST, Gaithersburg, Maryland, 2001.

[Smeulders *et al.*, 2000] Arnold W. M. Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22:1349–1380, December 2000.

[Smith and Chang, 1997] John R. Smith and Shih-Fu Chang. *Intelligent Multimedia Information Retrieval*, chapter Querying by Color Regions Using the VisualSEEK Content-Based Query System. California:AAAI Press/ The MIT Press, 1997.

[Spink *et al.*, 2002] Amanda Spink, Bernard J. Jansen, Dietmar Wolfram, and Tefko Saracevic. From e-sex to e-commerce: Web search changes. *IEEE Computer*, 35(3):107–109, 2002.

[Squire *et al.*, 1998] David McG. Squire, Wolfgang Müller, Henning Müller, and Jilali Raki. Content-based query of image databases, inspirations from text retrieval: inverted files, frequency-based weights and relevance feedback. Technical Report 98.04, Computer Vision Group, Computing Centre, University of Geneva, Geneva, Switzerland, November 1998.

[Squire *et al.*, 1999] D. Squire, W. Müller, H. Müller, and J. Raki. Content-based query of image databases, inspirations from text retrieval: inverted files, frequency-based weights and relevance feedback. In *The 10th Scandinavian Conference on Image Analysis (SCIA'99), (Kangerlussuaq, Greenland), June 7-11*, 1999.

[Stricker and Orengo, 1995] Markus A. Stricker and Markus Orengo. Similarity of color images. In *Storage and Retrieval for Image and Video Databases (SPIE)*, pages 381–392, 1995.

[Sunayama *et al.*, 2004] Wataru Sunayama, Akiko Nagata, and Masahiko Yachida. Image clustering system on WWW using web texts. In *Proceedings of the Fourth International Conference on Hybrid Intelligent Systems (HIS'04)*, pages 230–235, 2004.

[Takala *et al.*, 2005] V. Takala, T. Ahonen, and M. Pietikäinen. Block-based methods for image retrieval using local binary patterns. 2005. In: Image Analysis, SCIA 2005 Proceedings, Lecture Notes in Computer Science 3540, Springer, 882-891.

[Tamura *et al.*, 1978] H. Tamura, S. Mori, and T. Yamawaki. Textual features corresponding to visual perception. *I.E.E.E. Transactions on Systems, Man, and Cybernetics*, SMC-8, 1978.

[Teevan *et al.*, 2004] Jaime Teevan, Christine Alvarado, Mark S. Ackerman, and David R. Karger. The perfect search engine is not enough: A study of orienteering behavior in directed search. In *CHI '04: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 415–422, New York, NY, USA, 2004. ACM Press.

[Tsikrika and Westerveld, 2008] Theodora Tsikrika and Thijs Westerveld. Focused access to XML documents. chapter The INEX 2007 Multimedia Track, pages 440–453. Springer-Verlag, Berlin, Heidelberg, 2008.

[van Zwol *et al.*, 2008] Roelof van Zwol, Vanessa Murdock, Lluis Garcia Pueyo, and Georgina Ramirez. Diversifying image search with user generated content. In *Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval*, MIR '08, pages 67–74, New York, NY, USA, 2008. ACM.

[von Ahn and Dabbish, 2004] Luis von Ahn and Laura Dabbish. Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, CHI '04, pages 319–326, New York, NY, USA, 2004. ACM.

[Voorhees, 1993] Ellen M. Voorhees. Using wordnet to disambiguate word senses for text retrieval. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '93, pages 171–180, New York, NY, USA, 1993. ACM.

[Voorhees, 1994] E.M. Voorhees. Query expansion using lexical-semantic relations. In W. Bruce Croft and C. J. van Rijsbergen, editors, *Proceedings of the 17th Annual International ACM SIGIR Conference on*

*Research and Development in Information Retrieval (SIGIR '94), July 3-6, 1994, Dublin, Ireland*, pages 61–69. Springer New York Inc., NY, USA, 1994.

[Wang *et al.*, 1997] Jia Wang, Wen-jann Yang, and Raj Acharya. Color clustering techniques for color-content-based image retrieval from image databases. In *Proceedings of the 1997 International Conference on Multimedia Computing and Systems*, pages 442–449, Washington, DC, USA, 1997. IEEE Computer Society.

[Wang *et al.*, 2008] Changhu Wang, Lei Zhang, and Hong-Jiang Zhang. Learning to reduce the semantic gap in web image retrieval and annotation. In *SIGIR '08: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 355–362, New York, NY, USA, 2008. ACM.

[Westerveld and de Vries, 2003a] Thijs Westerveld and Arjen P. de Vries. Experimental evaluation of a generative probabilistic image retrieval model on 'easy' data. In *Proceedings of the SIGIR Multimedia Information Retrieval Workshop*, 2003.

[Westerveld and de Vries, 2003b] Thijs Westerveld and Arjen P. de Vries. Experimental result analysis for a generative probabilistic image retrieval model. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, SIGIR '03, pages 135–142, New York, NY, USA, 2003. ACM.

[Westerveld *et al.*, 2003] T. Westerveld, A. P. De Vries, A. Van Ballegooij, F. De Jong, and D. Hiemstra. A probabilistic multimedia retrieval model and its evaluation. *Eurasip J. Appl. Sign. Process.*, 2003(2):186–198, 2003.

[Yahoo, 2006] Yahoo. Yahoo search engine, 2006. [Online; accessed 14-April-2006].

[Yin *et al.*, 2003] Xiaoxin Yin, Mingjing Li, Lei Zhang, and Hongjiang Zhang. Semantic image clustering using relevance feedback. In *IEEE International Symposium on Circuits and Systems*, 2003.

[Zamir and Etzioni, 1999] Oren Zamir and Oren Etzioni. Grouper: A dynamic clustering interface to Web search results. *Computer Networks*, 31(11–16):1361–1374, 1999.

[Zhao and Grosk, 2002] R. Zhao and W. Grosk. Narrowing the semantic gap - improved text-based web document retrieval using visual feature. *IEEE Transactions on Multimedia*, 4:189–200, June 2002.

[Zheng *et al.*, 2004] Xin Zheng, Deng Cai, Xiaofei He, Wei-Ying Ma, and Xueyin Lin. Locality preserving clustering for image database. In *MULTIMEDIA '04: Proceedings of the 12th Annual ACM International Conference on Multimedia*, pages 885–891, New York, NY, USA, 2004. ACM Press.