TELECOM ParisTech at ImageClefphoto 2008: Bi-Modal Text and Image Retrieval with Diversity Enhancement

Marin Ferecatu[*†] Hichem Sahbi[†*]
[*] Institut TELECOM, TELECOM ParisTech
[†] CNRS LTCI, UMR 5141
46, rue Barrault, 75634 Paris Cedex, France
Marin.Ferecatu@telecom-paristech.fr
Hichem.Sahbi@telecom-paristech.fr

Abstract

In this paper we describe the participation of TELECOM ParisTech in the ImageClefphoto 2008 challenge. This edition focuses on promoting diversity in the results produced by the retrieval systems. Given the high level semantic content of the topics, search engines based solely on text or visual descriptors are unlikely to offer satisfactory results. Our system uses several text and visual descriptors, as well as several combination algorithms to improve the overall retrieval performance. The text part includes a collection of manually built boolean queries and a set of textual descriptors extracted automatically using dictionary filtering and dimensionality reduction. Text and visual descriptors are combined using two strategies: ad-hoc concatenation and re-ranking. Diversity makes it possible to reduce the redundancy in the final results and it is obtained using two techniques, threshold clustering and maxmin exploration. Several runs were submitted to the challenge, including individual (text or visual), combined, and with different settings of diversity. The results show that the combined runs outperform by a significant amount the individual runs. These results clearly corroborate (i) the complementarity of text and visual descriptors and (ii) the effectiveness of boolean queries suggesting promising future research directions.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.3.7 Digital Libraries; H.2.3 [**Database Management**]: Languages—*Query Languages*

General Terms

Measurement, Performance, Experimentation.

Keywords

Image retrieval, Reranking, Support Vector Machines, Hybrid Text and Image Search.

1 Introduction

Stimulated by the exponential growth of multimedia contents, the interest in document indexing and retrieval has steadily increased in recent years. Although text based retrieval has been studied for several

decades [26], many problems remain unsolved [20]. Recently, text based retrieval emerged successfully in internet search [17], and it usually relies on statistical text processing and tags collections. Content based image retrieval also developed rapidly in the last decade, motivated by the growing amount of image and video collections and by the need to organize, share and search those contents effectively and efficiently [28, 4]. Since neither world (text or image indexing and retrieval) offers a satisfactory bridge to solve the infamous semantic gap, more and more search engines employ hybrid text and visual representations in order to describe and search multimedia databases. In this context, the ImageClef challenge offers an excellent benchmark to test and compare several state of the art algorithms proposed by different participants[1].

ImageClefphoto 2008 focuses on promoting diversity in the results produced by search engines, i.e., those which reduce the number of redundant elements in their output are preferred. As we shall see in §2, the query topics are very semantic so individual text or visual descriptors are not expected to offer satisfactory results. In this work, we investigate several combination strategies for text and visual descriptors as well as several algorithms for reducing the redundancy in the results returned by the system. In order to describe the textual content we use two representations: (1) a set of manually constructed boolean queries and (2) a set of automatically extracted vector representations based on dictionary filtering and dimensionality reduction. We also use several global image descriptors [19, 4], which even though less performant than local ones [24], have the advantage of being generic and computationally cheap (see §3).

We compare hybrid combination by concatenation and re-ranking, using both the Query By Example (QBE) paradigm and SVM learning. In order to eliminate the redundancy in the final results we employ two strategies: a modified version of Quality Thresholding algorithm and a maxmin exploration strategy. Our results show that combining the visual search results (even when using a small number of examples, e.g. three in the challenge) with the textual results improves significantly the overall performance. Clearly, the information provided by the two modalities are complementary. Furthermore, the manually prepared boolean queries provide a noticeable and consistent gain suggesting that semantic parsing and automatic extraction of boolean representations is a promising research direction.

The paper is organized as follows. We start by a short presentation of the ImageClef Photo Retrieval Task (§2). Then, we describe our visual descriptors and the retrieval results obtained using image descriptors only (§3). Manual queries are described in §4.1 while in §4.2 we present our automatic text feature extraction from raw data and their combination with the visual descriptors. In §5 we describe our diversification algorithm; we end the paper with a discussion and concluding remarks.

2 ImageCLEF Photo Retrieval Task

In this section we briefly describe the ImageClefphoto retrieval challenge, description that we use later to motivate our choices and to discuss various results. This year, ImageClefphoto focuses on promoting diversity in the results produced by search engines, i.e. those which eliminate (near-)duplicate documents are likely to produce a higher percentage of meaningful results. The performance measures used in the challenge are the precision at 20 (P20) defined as

$$P20 = \frac{\text{Relevant}(20)}{20} \tag{1}$$

and the cluster-recall at 20 (CR20), or sub-topic recall:

$$CR20 = \frac{\left|\bigcup_{i=1}^{20} S(d_i)\right|}{N_S} \tag{2}$$

where for a given topic, the document d_i is relevant for the set of sub-topics $S(d_i)$ and N_S is the total number of sub-topics [31]. The submitted runs were ranked according to the average of the two above measures.

The benchmark uses the IAPR-TC12 [12] data collection, which consists of 20,000 images associated with a set of XML annotations, including text descriptions, location tags, title and date. The challenge proposes 39 query topics, also available in XML format. Each topic is defined by a narrative block indicating

[1] 25 research teams submitted runs to the ImageClefphoto 2008 challenge

<DOC>
<DOCNO>annotations/00/60.eng</DOCNO>
<TITLE>Palma </TITLE>
<DESCRIPTION>two lane street with large shops on the right and smaller shops on the left; people are walking on the sidewalk, some are crossing the street; cars are parked along the left side of the street as well; </DESCRIPTION>
<NOTES>The main shopping street in Paraguay; </NOTES>
<LOCATION>Asunción, Paraguay </LOCATION>
<DATE>March 2002 </DATE>
<IMAGE>images/00/60.jpg </IMAGE>
<THUMBNAIL>thumbnails/00/60.jpg </THUMBNAIL>
</DOC>

<top>
<num> Number: 5 </num>
<title> animals swimming </title>
<cluster>animal</cluster>
<narr> </narr>
<image> 3739.jpg </image>
<image> 4968.jpg </image>
<image> 30823.jpg </image>
</top>

Figure 1: IAPR-TC12 database entry description (left) and query topic format (right).

the search target, a diversity criterion and a set of three image queries (Fig. 1 shows one topic description and a database entry). A short analysis of these data reveals that the topics combine highly semantic concepts and complex logical structures. For each topic, three images are also supplied in order to enhance understanding and to help formulating visual queries, but they are not intended as a replacement of text descriptions. Indeed, due to the complex semantic definitions of the topics, image queries are unlikely to retrieve all relevant results. Thus, image examples are perhaps best used to enhance the results through their combination with the text description, and this is the approach we adopted for this challenge.

Different acronyms were used by participants and stand for the following runs and query types: IMG (image), TXT (text) and TXTIMG (combination text/image), MAN (manual) and AUTO (automatic).

3 Visual Content Description and Search

In this section we motivate our choice of image descriptors for the ImageClefphoto retrieval use case (see §3.1 and §3.2). We then present, in §3.3, our querying paradigms and we show in §3.4 some performance measurements, including a comparison with other ImageClef runs. We end this section with a brief discussion of the limits of visual queries.

3.1 Motivation

Extracting visual features that capture high level semantic content of an image is a difficult task. Indeed, the main challenge for content based retrieval systems is the infamous "semantic gap" that instantiates the discrepancy between the low level features used to represent the visual content and the high level concepts expected by the users [4].

For some domain specific databases and applications, such as browsing fingerprint or face images, there exists enough *a priori* knowledge of the image content to be able to propose more accurate mathematical models. However, for generic databases, the task becomes much harder since there is no perfect and unique description of the visual content which agrees with the semantic[2]. Therefore, many systems rely on *holistic* combined image descriptions such as color, texture and shape [4, 18, 10] in order to search for target images. This approach, even though *ad-hoc*, has been successfully applied for unstructured image databases (which lack text descriptions) through an interactive description of concepts using relevance feedback and machine learning techniques [33].

Recent object recognition approaches rely on *local* descriptors (for example, image regions or interest points) in order to describe more precisely the visual content [24]. While local descriptors have many desirable properties, such as stability and invariance to common geometric and photometric transformations [23, 32], they are resource (time and memory) demanding and cannot be easily extended to large-scale search engines. Moreover, although these algorithms can be tuned to perform well for some object cate-

[2]This is due to lack of consensus about the underlying semantics, which usually depends on the context (even humans disagree when interpreting images.)

gories, they are less adapted to pictures involving deformable objects or context dependent scenes that are difficult to describe using individual rigid objects (for example, emotional states or esthetic impressions.)

Since ImageClefphoto query topics are very semantic and sometimes expressed as a combination of several concepts, they are not easily translated in terms of low level visual features. The large number of topics and concepts involved in their definition rules out the possibility to use specific visual models for each concept. Instead, we use global image descriptors as described below. While being less performant compared to local descriptions, they are more appropriate for our use case as (1) they have small memory footprint and thus fit into standard PCs without any specific storage requirements; and (2) they are very fast to compute as they involve simple distance measure operations, guaranteeing real time responses. Furthermore, as they do not include any a priori object model, they can be applied to any target category. Indeed, global descriptors have been shown to perform well in this framework, for example with SVM-based machine learning [33].

3.2 Global Image Descriptors

As described in §3.1, we use global image descriptors in order to represent the visual content of images. More precisely, we use a combination of color, texture and shape features, as described below.

Color histograms: they provide a summary description of the color information but ignore spatial correlations between colors; thus, pixels having the same color distribution may not be similar in the context of their spatial neighborhoods [30], [1]. Alternatively, given an image of size $M \times N$, we weight each color instance by a measure related to its local context:

$$h(\mathbf{c}) = \frac{1}{MN} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} w(x,y)\delta(f(x,y) - \mathbf{c})$$

where $h(\mathbf{c})$ is frequency of color $\mathbf{c}$ and $w(x,y)$ is a pixel-based weighting function. We use $w(x,y) = ||\Delta(x,y)||^2$, the Laplacian at the pixel (x,y), to emphasize corners and edges in the image and local color frequency to emphasize non-uniform regions.

Texture features: we use the power spectral density distribution in the complex plane. This has been shown to perform well when combined with color and shape histograms [19]. Roughly, a high energy spectrum concentrated at low frequencies highlights large scale informations in an image, while high frequencies correspond to textured regions (small scale details).

Shape features: in order to describe the shape content of an image we use standard edge orientation histograms. First, edges are extracted from images, then the gradient is computed using only the edge pixels. The orientation of the gradient is quantized w.r. to the angle resulting into a histogram that is sensible to the general flow of lines in the image [14]. More details on image descriptors can be found in [7].

Visual feature vectors are combined by concatenation and then, in order to reduce resource requirements and to avoid the curse of dimensionality, we apply linear PCA [16] and keep the 100 largest principal components (which preserves 95% of the energy of the signal.)

3.3 Querying Paradigms

For each topic, we use the three query images combined with two different paradigms: (i) similarity search by minimum distance and (ii) SVM filtering.

MinDistance Query: let $\mathcal{B}$ be the database and let $Q = \{q_1, q_2, q_3\}$ denote the query, here q_1, q_2 and q_3 are the three images of the query topic. We extend the "query by example" search paradigm for multiple inputs by introducing a composed measure of dissimilarity between an image $x \in \mathcal{B}$ and Q:

$$d(x, Q) = \min_{i=1,2,3} d(x, q_i), \tag{3}$$

This definition naturally follows from the fact that images in $\mathcal{B}$ close to one of the three query images $\{q_i\}$ are more likely to belong to the same topic. Instead of taking the $\min$-value, one can consider a

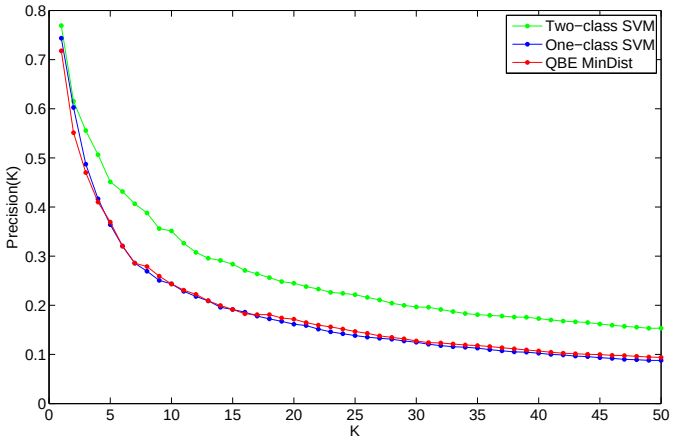

Figure 2: Comparative performance of two-class SVM, one-class SVM and similarity retrieval with MinDist.

convex combination of distances; nevertheless, if $\{q_i\}$ are very distant in the description space, the resulting dissimilarity measure is uncorrelated with the semantic captured by the query topics.

SVM Based Querying: Support Vector Machines (SVM) [27] have been applied with success to a great wealth of practical problems since their inception in early '90s by Vapnik [29]. In image retrieval, they provide state of the art results in many recognition tasks [4] and relevance feedback [33]. They use a linear combination of kernels as a decision function:

$$f_{SVM}(x) = \sum_{i=1}^{N} \alpha_i K(x, x_i), \tag{4}$$

where $K(\cdot,\cdot)$ is a positive definite kernel, $\{\alpha_i\}$ are the signed Lagrange multipliers and $\{x_i \ : \ \alpha_i \neq 0\}$ are known as the support vectors (see [27, 29] for details).

We trained our SVMs using $Q = \{q_1, q_2, q_3\}$ as positive examples. One may use one class SVMs for training, but their generalization performances were reported to be sub-optimal (with respect to standard SVMs) for image retrieval [4, 6] mainly when the positive classes contain few examples. In practice, we used, for each topic, standard two class SVMs trained on Q and random subsets of 10 negative examples taken from $\mathcal{B}$. As the size of $\mathcal{B}$ is $20,000$, it is unlikely that some relevant images belong to these negative random subsets. While simple, this procedure proved to be effective in practice and produced better results compared to one class SVMs.

Once each topic SVM learnt[3], the system ranks the database according to the score given by Eq. 4 and returns the most positive examples. In all these experiments, we use the Laplacian kernel defined as $K(x_i, x_j) = \exp\left(-\gamma \|x_i - x_j\|\right)$. This kernel was advocated in [3] for histogram-based image description and proved to provide better results than the usual Gaussian kernel for relevance feedback tasks [15]. Notice that the dominant term in its Taylor expansion corresponds to the triangular kernel, $K(x_i, x_j) = -\gamma \|x_i - x_j\|$, which is proved to be scale invariant with respect to the distribution of the data in the description space [8]. In practice, we found that a good setting of γ is 1. For consistency and comparison issues, we fix in all our experiments the kernel and its parameters. Of course, our results could be improved by fine tuning the kernel parameters or by exploring other kernels, but this is not in the scope of this work.

3.4 Performance

In Fig. 2 we present the average precision over all topics as a function of the number of results sent by the query engine. At this stage, we draw the following conclusions:

[3] In our experiments we used the well known LIBSVM package [2], see `http://www.csie.ntu.edu.tw/~cjlin/libsvm`

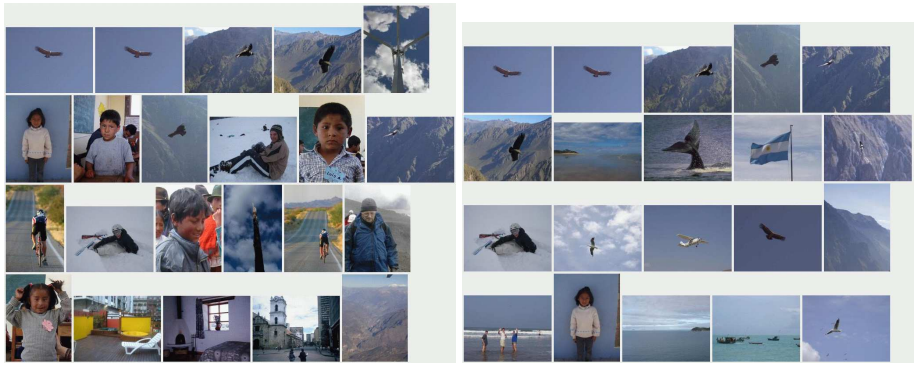

Figure 3: "Bird flying" search topic: comparative results for similarity retrieval by MinDist (left) and SVM retrieval (right).

1. As expected, the SVM query procedure described above outperforms all the other paradigms (a gain of almost 10% when compared to QBE MinDist and One-class SVM).

2. One-class SVM performs about the same as the QBE MinDist. This can be explained by the fact that the number of learning examples is very small (only three) and thus, not enough to capture the complexity of the target topic.

These runs do not include diversity and were not submitted to the ImageCLEF-PRT (see §5). The two IMG runs we submitted (using only visual features), were ranked 2^{nd} and 3^{rd} tailing the 1^{st} rank run performances on the combined P20/CR20 measure (see §5). In terms of the $P20$ measure (Eq 1), our two runs were ranked 4^{th} and 6^{th} which proves that the diversification algorithms, although lowering the $P20$ measure, improved the overall performance. As an illustration, Fig 3 shows the difference between SVM and direct MinDist similarity retrieval using the "birds flying" topic. As expected, SVM uniformly provides more consistent results.

3.5 Limits of the Visual Search

Motivated by a wealth of practical applications, image retrieval by visual content has became a rapidly evolving research field, although breakthrough advances are still rare. State of the art results are far from satisfactory and search by image descriptors alone is unlikely to offer complete satisfaction for most practical task [4]. This state of progress clearly motivates the use of hybrid descriptions, for example by combining visual features with text and other available media (sound, music, meta-tags, etc.) Meanwhile, recent trends in research suggest that machine learning based search methods with relevance feedback provide excellent results for many tasks. In the following sections we describe our approach for ImageClefphoto challenge by using manually prepared boolean queries (§4.1) and combined text and image content representations (§4.2).

4 Hybrid Document Search

As mentioned in §3, using only visual descriptors is not enough to provide satisfactory results in this challenge. Our goal is to measure the gain when combining text and image features; in §4.1 we describe boolean queries and their combination with visual descriptors while in §4.2 we present our text descriptor based on dictionary filtering and dimensionality reduction. Both approaches are illustrated by examples of actual queries; we also present different results from the challenge.

Figure 4: Some examples of boolean retrieval (see the text for details).

4.1 Boolean Queries

A short analysis of the query topics reveals complex and highly semantic concepts involving several objects and relations. One possible way to represent these topics in a principled way is to use boolean queries. Let us consider the 3rd topic ("religious statue in the foreground"):

> 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.

This can naturally be expressed using boolean operations involving concepts and operations: "(religious AND statue) AND NOT (memorial OR monument OR war) AND NOT (LOCATION 'Easter Island')". Boolean retrieval relies on the use of logical operators where the terms in queries are linked together using an algebra of simple operations (including AND, OR and NOT). Nevertheless, automatic extraction of boolean queries from raw text is known to be a difficult and still unsolved task [11, 20]. Hence, we choose to manually build these expressions from the raw text. This approach which is very popular in Internet search, has emerged as one of the easy to use standards in text retrieval.

4.1.1 Query Construction

We first introduce a small querying language adapted for the ImageClefphoto challenge. We use AND, OR and NOT to connect terms and we use a "LOCATION" specifier to makes it possible to filter documents by their locations (such as "country" or "city" tags). These informations can easily be extracted from the XML document descriptions.

For some topics, we filter and tag (as "BW": black and white) images depending on their grey level information. This is implemented using the saturation component in the HSV color space. For instance, the query "church AND BW AND NOT LOCATION France" seeks grey level pictures of churches not located in France.

Queries are created using a web interface. Using the raw text, the user interactively formulates boolean queries for different topics. This procedure is similar to relevance feedback as the user iteratively updates the boolean queries until the results are satisfactory. Fig. 4 shows some results: (left) topic #2 "((church OR cathedral OR mosque) AND (towers) AND (three OR four OR five))", (middle) topic #11 "((LOCATION Russia) AND (BW))" and (right) topic #2 "(lighthouse AND (water OR sea OR ocean))".

4.1.2 Combination with Visual Descriptors

Boolean queries described earlier return a set of candidate relevant images which are not scored (and hence unranked) so it is not possible, at this stage, to use merging techniques in order to produce combined text/image ranks. Instead, we intersect the results of image and text queries. Notice that topics are very complex and difficult to express exactly using only boolean queries. The latter may produce relevant and (also) irrelevant results which are filtered using visual search[4].

[4] In practice, visual queries return 1000 documents which are intersected with the results of text and scored using visual distances.

4.1.3 Experiments

In our experiments, we extract the raw text, the LOCATION field from XML documents and the "BW" tag from the underlying images. Using these informations and the boolean queries, the search engine returns the results in real time. Notice again that the user is involved only in the construction of boolean queries so all further steps, are *fully automatic*. Fig. 5 shows the evaluation of different querying schemes (text, image and combined text/image).

We clearly see that boolean queries outperform visual searches. This is predictable since the former are manually and interactively constructed while visual search is fully automatic. We also see from the results that combining visual and boolean searches boosts the precision by $\sim 20\%$ and even though limited, the information provided by the visual descriptors is always informative. This result is slightly unexpected, because of the small number of positive examples (only three) and suggests that visual descriptors provide a complementary information w.r. to the boolean queries. We noticed that the influence of the visual descriptors decreases as the rank increases: at rank 50, combined text/image performs about the same as text only. This is due to the small number of examples available to compute the visual similarity, e.g. for lower ranks the results are less semantically correlated with the topics. Finally, we conclude that hybrid search always outperforms individual text or image searches.

The runs we submitted to the ImageClefphoto were similar, but they differ in terms of their diversity algorithms (see. §5) and were ranked 1st, 2nd and 3rd in ImageClefphoto.

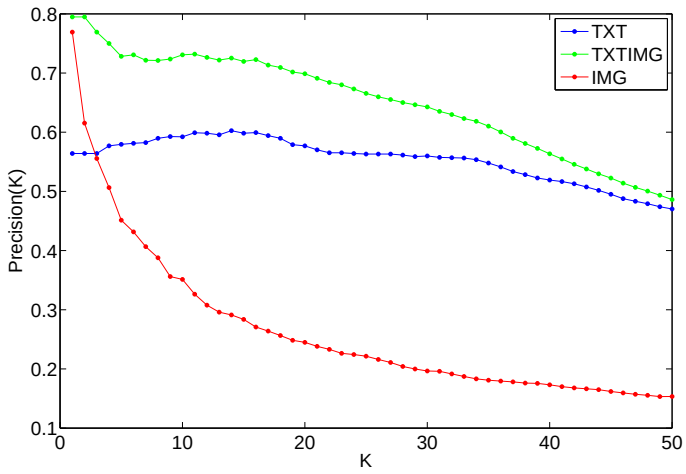

Figure 5: Precision of boolean retrieval: text and image alone versus combined text and visual queries.

4.2 Automatic Queries

In this section, we compare several schemas that combine automatically extracted text descriptors and visual features. The automatic query description and querying is expected to be less performant than the manual one. As described below, we consider in these experiments both early and late descriptor merging techniques.

4.2.1 Text Description

Text indexing and retrieval is a growing field and many existing state of the art techniques provide reasonably accurate results [26, 22, 17]. Nevertheless, most of them rely on the estimation of statistical measures and can only be applied on large datasets. For the IAPR-TC12 database, almost all the terms appear only once, so these statistical techniques are clearly not applicable. In order to extract meaningful information from short descriptions, semantic analysis and natural language parsing are necessary; however, these are known to be difficult and still unsolved problems [21, 20].

Our goal is to investigate the performance improvement obtained by combining both textual and visual descriptions, so we adopted a simple vector space model in order to represent the text information. First, we eliminate the stop words, then we parse the list of terms with the Porter stemmer[5][25]. We associate to each resulting term a coordinate in the feature space. As no relationships are considered between terms, we only keep the terms that are used in the definition of the query topics. A further step is considered in order to reduce the dimensionality based on a linear version of PCA. In this step, we only keep the first 100 principal components, corresponding to 98% of the total statistical variance of the text data. We measure dissimilarity between documents using the L1 distance (cosine distance produced similar results). Query topics are described in a similar way but they are pre-process ed in order to eliminate words common to all topics[6].

4.2.2 Combination with Visual Descriptors

We use two merging strategies: 'Early merging' refers to combining descriptors prior to querying while 'late merging' performs individual text and visual queries prior to combination.
Early merging: we create hybrid descriptors by concatenation (cartesian product space) of text and the visual feature spaces. To ensure equal contributions in the final feature vector, we normalize individual features by the mean and the variance computed on the whole dataset.
Late merging: for each document, we first run individual queries (text and visual) obtaining two ranking lists. Then each document is assigned a final rank based on the MINRANK scheme defined as

$$r(I) = \min(r_V(I), r_T(I)) \tag{5}$$

where $i \in \mathcal{B}$ is an image from the database $\mathcal{B}$ and r_V (resp. r_T) is its visual (resp. text) rank. The intuition behind this combination strategy comes from the fact that the best rank should be preferred, e.g. low ranked documents are unlikely to be similar to the query topic.

4.2.3 Experiments

Fig. 6 shows a comparison of the merging techniques. As we see, textual features extracted automatically outperform significantly the visual ones. We also notice that early combination by concatenation (cartesian product) produces slightly better results than the MINRANK late combination, but the difference is not significant. From these experiments, the best results were obtained by applying the MINRANK algorithm (as described previously) to the hybrid text/image and text only. We obtain $\sim 10\%$ improvement w.r.t. to text retrieval. Nevertheless, this gain is less significant compared to boolean queries which are built manually (see §4.1).

Finally, this run is ranked 6$^{\text{th}}$ in the ImageClefphoto AUTO_TXTIMG and even though no diversification algorithm was used, it ranked 2$^{\text{nd}}$ w.r. to the CR20 measure. This can be explained by the fact that the MINRANK mixes best ranks from both image and text results and since text and visual features are independently extracted, some reduction in the redundancy of the results is expected, i.e. the returned images do not belong to the same sub-categories.

5 Diversity and Clustering

As presented in §2, the measures used to evaluate the runs are precision and cluster-recall on the first 20 results (denoted as P20 and CR20 resp.) According to these measures, results should be both relevant and as diverse as possible. Notice that extra information is also provided for each query topic (specified by the field "`<cluster></cluster>`") about the targeted diversity criterion. In this section, we describe our diversity schemes and their impacts on the P20 and CR20 performance.

[5] `http://tartarus.org/~martin/PorterStemmer`
[6] For example, expressions like "relevant images" are used in all topic descriptions.

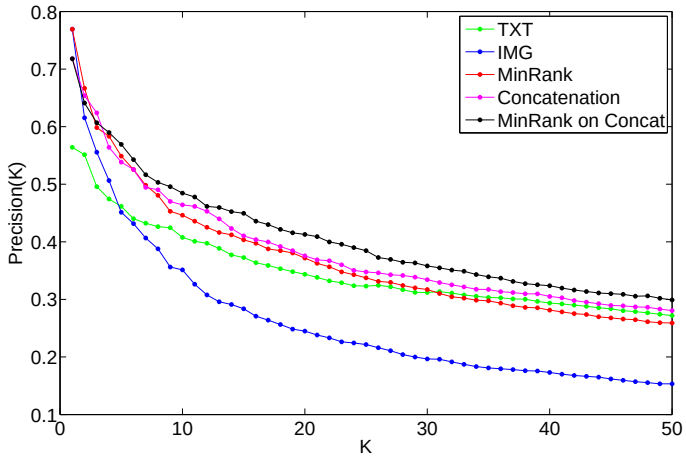

Figure 6: Precision results for several text and image combination techniques.

5.1 Text Clustering

Some of the diversity criteria required by the query topics can be directly extracted from the XML data. Each document contains a tag such as "city", "state" and "location" which can be used in order to filter and group documents. For that purpose and in order to keep a good balance between precision and diversity, we start with a larger selection of 40 results and we cluster them following the specified location tags. The output is formed by taking the 1st element in each cluster, then the 2nd, etc.

5.2 Visual Diversification

Visual clustering is applied when the diversity criterion cannot be extracted from the XML tags associated to images. We consider two types of "visual diversity" algorithms, described below.

5.2.1 MAXMIN Diversity

The MAXMIN diversity algorithm is based on the maximization of the mininum distance of a given document with respect to the (so far) selected results. Our algorithm starts by choosing the document with the best rank. Afterwards, it chooses the next document as the one which maximizes the distance with respect to the first. More precisely, let $\mathcal{S}$ be the candidate set (the initial window of size 40 in our case) and suppose that $\mathcal{C} \subset \mathcal{S}$ is the set of already selected examples. Then, the next example is chosen as:

$$x = \text{argmax}_{x_k \in \mathcal{S} \setminus \mathcal{C}} \min_{x_i \in \mathcal{C}} d(x_k, x_i) \tag{6}$$

This procedure does not produce a clustering, but rather a permutation of the initial selection such as its prefixes are very diversified according to Eq. 6 with respect to the distance defined on the description space.

5.2.2 QT Clustering

Our second "diversification" method is based on visual clustering. We tested several standard clustering techniques, such as Fuzzy-K-Means [5] and Competitive Agglomeration [9], but we obtained many clusters of large size and highly diverse semantic content. To control the size of the generated clusters, we developed a variant of the Quality Threshold algorithm [13], described below.

Let s denotes the cluster size defined as $s = N/n_C$, where N, n_C are respectively the number of images and the expected (fixed) number of clusters. The algorithm builds the clusters iteratively. First the center of the new cluster is chosen by minimizing the following criterion

$$x_t = \text{argmin}_{x_i \in S_t} \mathcal{R}\left(\text{KNN}_s(x_i; \mathcal{S}_t)\right) \tag{7}$$

where $\mathrm{KNN}_s(x_i;\mathcal{S}_t)$ denotes the set of s nearest neighbors of x_i in $\mathcal{S}_t$ ($\mathcal{S}_1 = \{x_1, ..., x_N\}$) and $\mathcal{R}$ the radius of the smallest sphere enclosing $\mathrm{KNN}_s(x_i;\mathcal{S}_t)$. The new cluster is built as $C_t = \mathrm{KNN}_s(x_i;\mathcal{S}_t)$, and then removed from the remaining data, i.e. $\mathcal{S}_{t+1} = \mathcal{S}_t \backslash \mathcal{C}_t$. As for the text clustering, the final output is formed by taking the 1st element in each cluster, then the 2nd, etc., until all elements are exhausted.

5.3 Experiments

We submitted several runs to ImageClefphoto including the diversity settings described above. More specifically, 26 out of the 39 topics have "cluster tags" related to the location of the pictures; diversity is performed for these topics using text clustering while for the rest it is visual. We observed that applying a diversity schema lowers the P20 measure; however, as the CR20 measure increases, we expect the results to be more meaningful.

Visual Retrieval allows us to evaluate the impact of visual diversity (using QT or MAXMIN clustering). As already discussed in §3.4, the best "visual retrieval" performances were achieved using the two class SVM (P20=0.248) and without diversity. We submitted two runs to the challenge using QT and MAXMIN clustering and they are ranked 2nd and 3rd when using the combined P20 and CR20 score. Their P20 measure is, as expected, lower (0.20 and 0.17 respectively) when compared to the non-diversified results, and this suggests that diversity indeed produced an important CR20 gain.

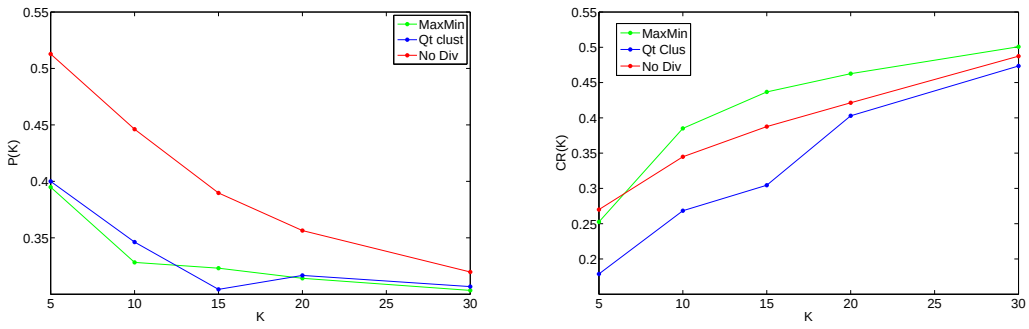

Figure 7: Hybrid text and image descriptor: the effect of "diversification" of results.

Combined Text/Visual Search: in Fig. 7 we examine the impact of different diversity algorithms on the P20 and CR20 measures for the combined text and image descriptors (see §4.2.2). First, we observe a loss of precision for both QT and MAXMIN algorithms. Nevertheless, the QT algorithm failed to produce better CR rankings when compared to the case without diversity. For this algorithm, we set the number of clusters to 20 and the initial selection contains the 40 first ranked results. As the size of different clusters is very small (i.e., $s = 2$), many clusters are similar and this affects the quality of diversity. Moreover, even for a perfect retrieval, there is simply not enough data (less than 100 in average) per class in order to generate consistent clustering. These results show that the space of query results is better explored and summarized using MAXMIN than QT clustering which suffers from the insufficient amount of data.

6 Conclusion and Perspectives

In this paper we presented our experiments investigating the performance of several combination techniques for image and text descriptors, on the ImageClefphoto challenge database. We compared early merging of descriptors by concatenation with late ranking combination obtained by separate queries on text and image features. We also described two schemes for reducing redundancy in the results returned by our search engine.

In our first conclusion, we found that even with very few images (three in the ImageClefphoto), our system was able to improve the results significantly. Moreover, the improvement is more significant in case of manually prepared boolean queries. This clearly indicates that good quality boolean queries are less

likely to return noisy results with respect to the targeted topic. Automatic extraction of boolean queries from raw text is hence identified as a worthy to explore research direction, for instance by using Parts of Speech (POS) tagging and language parsing.

In our second conclusion, we noticed that using a diversification algorithm helped improving the ranking of our submitted runs. This is more noticeable for queries using only visual descriptors (see §3) where the proposed diversification schemes significantly improved the ranking of our runs (2nd and 3rd). However, because of the limited size of ground truth classes (less than 100 images per topic), it is not possible, at this stage, to draw firm conclusions. Indeed, in a real search engine, where topics might be represented by millions of (possibly similar) images, we expect the obtained clusters to be much more consistent.

Acknowledgements

This work was supported by the French National Research Agency (ANR) under the AVEIR[7] project, ANR-06-MDCA-002.

References

[1] Nozha Boujemaa, Julien Fauqueur, Marin Ferecatu, François Fleuret, Valérie Gouet, Bertrand Le Saux, and Hichem Sahbi. Ikona: Interactive generic and specific image retrieval. In *Proceedings of the International workshop on Multimedia Content-Based Indexing and Retrieval (MMCBIR'2001)*, 2001.

[2] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: a library for support vector machines. Technical report, National Taiwan University, 2001. Software available at http://www.csie.ntu.edu.tw/˜cjlin/libsvm.

[3] Olivier Chapelle, P. Haffner, and Vladimir N. Vapnik. Support-vector machines for histogram-based image classification. *IEEE Transactions on Neural Networks*, 10(5):1055–1064, 1999.

[4] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys*, 40(2):5:1–60, 2008.

[5] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*. Wiley Interscience, 2001.

[6] M. Ferecatu, N. Boujemaa, and M. Crucianu. Semantic interactive image retrieval combining visual and conceptual content description. *ACM Multimedia Systems Journal*, 13(5–6):309–322, 2008.

[7] Marin Ferecatu. *Image retrieval with active relevance feedback using both visual and keyword-based descriptors*. PhD thesis, INRIA—University of Versailles Saint Quentin-en-Yvelines, France, 2005.

[8] François Fleuret and Hichem Sahbi. Scale-invariance of support vector machines based on the triangular kernel. In *3rd International Workshop on Statistical and Computational Theories of Vision*, October 2003.

[9] Hichem Frigui and Raghu Krishnapuram. A robust competitive clustering algorithm with applications in computer vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(5):450–465, 1999.

[10] Theo Gevers and Arnold W. M. Smeulders. Content-based image retrieval: An overview. In G. Medioni and S. B. Kang, editors, *Emerging Topics in Computer Vision*. Prentice Hall, 2004.

[11] David A. Grossman and Ophir Frieder. *Information Retrieval: Algorithms and Heuristics*. Springer, 2004.

[7] `http://aveir.lip6.fr`

[12] M. Grubinger, P. Clough, H. Muller, and T. Deselaers. The iapr tc-12 benchmark: A new evaluation resource for visual information systems. In *In Proceedings of International Workshop OntoImage'2006 Language Resources for Content-Based Image Retrieval, held in conjuction with LREC'06*, 2006.

[13] L.J. Heyer, S. Kruglyak, and S. Yooseph. Exploring expression data: Identification and analysis of coexpressed genes. *Genome*, 9(11):1106–1115, 1999.

[14] A.K. Jain and A. Vailaya. Shape-based retrieval: a case study with trademark image databases. *Pattern Recognition*, 31(9):1369–1390, 1998.

[15] Feng Jing, Mingjing Li, Lei Zhang, Hong-Jiang Zhang, and Bo Zhang. Learning in region-based image retrieval. In *Proceedings of the IEEE International Symposium on Circuits and Systems*, 2003.

[16] I.T. Jolliffe. *Principal Component Analysis*. Springer-Verlag, 2002.

[17] Amy N. Langville and Carl D. Meyer. *Google's PageRank and Beyond: The Science of Search Engine Rankings*. Princeton University Press, 2006.

[18] M. Lew, N. Sebe, C. Djeraba, and R. Jain. Content-based multimedia information retrieval: State-of-the-art and challenges. *ACM Transactions on Multimedia Computing, Communication, and Applications*, 2(1):1–19, 2006.

[19] B.S. Manjunath, P. Salembier, and T. Sikora, editors. *Introduction to MPEG-7: Multimedia Content Description Interface*. Wiley, 2002.

[20] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schtze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[21] Christopher D. Manning and Hinrich Schuetze. *Foundations of Statistical Natural Language Processing*. The MIT Press, 1999.

[22] Charles T. Meadow, Bert R. Boyce, Donald H. Kraft, and Carol L Barry. *Text Information Retrieval Systems*. Academic Press, 2007.

[23] Krystian Mikolajczyk and Cordelia Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 27(10):1615–1630, 2005.

[24] Jean Ponce, Martial Hebert, Cordelia Schmid, and Andrew Zisserman. *Towards category-level object recognition*, volume 4170. Springer, 2006.

[25] M.F. Porter. An algorithm for suffix stripping, program. *Program*, 14(3):130–137, 1980.

[26] Gerard Salton and Michael J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1986.

[27] Bernhard Schölkopf and Alexander Smola. *Learning with Kernels*. MIT Press, 2002.

[28] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, 2000.

[29] V. N. Vapnik. *Statistical Learning Theory.* John Wiley, September 1998.

[30] Constantin Vertan and Nozha Boujemaa. Upgrading color distributions for image retrieval: can we do better? In *International Conference on Visual Information Systems (Visual2000)*, November 2000.

[31] C.X. Zhai, W.W. Cohen, and J. Lafferty. Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In *In Proceedings of the 26th Annual international ACM SIGIR Conference on Research and Development in informaion Retrieval*, August 2003.

[32] Jianguo Zhang, Marcin Marszałek, Svetlana Lazebnik, and Cordelia Schmid. Local features and kernels for classification of texture and object categories: a comprehensive study. *International Journal of Computer Vision*, 73(2):213–238, 2007.

[33] Xiang Sean Zhou and Thomas S. Huang. Relevance feedback for image retrieval: a comprehensive review. *Multimedia Systems*, 8(6):536–544, 2003.