Semantic interactive image retrieval combining visual and conceptual content description

Marin Ferecatu* Nozha Boujemaa Michel Crucianu

Abstract

We address the challenge of semantic gap reduction for image retrieval through an improved SVM-based active relevance feedback framework, together with a hybrid visual and conceptual content representation and retrieval. We introduce a new feature vector based on projecting the keywords associated to an image on a set of "key concepts" with the help of an external lexical database. We then put forward two improvements of SVM-based relevance feedback method. First, to optimize the transfer of information between the user and the system, we introduce a new active learning selection criterion that minimizes redundancy between the candidate images shown to the user. Second, as most image classes span a wide range of scales in the description space, we argue that the insensitivity of the SVM to the scale of the data is desirable in this context and we show how to obtain it by using specific kernel functions. Experimental evaluations show that the joint use of the new concept-based feature vector and the visual features with our relevance feedback scheme may significantly improve the quality of the results.

1 Problem statement

The amount of multimedia documents has steadily increased in the recent years and with it the need for efficient organization and retrieval of these contents. Early on, thanks mostly to its easiness, keyword-based search became the leading paradigm for querying multimedia databases. One consequence was to stimulate professionals in continuously developing annotations for their multimedia databases. However, the use of keywords alone raises several important problems: the manual annotation is very expensive and inherently incomplete, the relation between words and concepts is often complex due to such phenomena as synonymy (different words denote the same concept) or homonymy (same word denotes different concepts) and many search criteria just can't

*INRIA Rocquencourt, IMEDIA Research Project, BP 105 Rocquencourt, 78153 Le Chesnay Cedex - France
Email: Marin.Ferecatu@inria.fr

be well described by a few keywords. These important difficulties have boosted research activities in the field of content-based image retrieval (CBIR) [13], [36]. However, the use of the visual content has its own limitations, due mainly to the semantic gap, which express the discrepancy between the low-level features that can be readily extracted from the images and the high level descriptions which are meaningful to the users.

In this paper, we explore the important issue of semantic gap reduction through **an unified twofold approach** integrating cross-media text and visual content representation with an improved relevance feedback framework. **First**, we argue that keywords and visual features are complementary descriptions and using both of them to query image databases offers more meaningful results to the users. Indeed, keywords offer a high level description of the image content and context, while the visual features bring a low level description which is often difficult to express in words. We introduce a new feature vector based on projecting the keywords associated to an image on a set of "key concepts" chosen with the aid of an external ontology. Our feature vector takes into account different semantic relations between keywords and may help finding documents with similar meaning even if they are annotated with different keywords. **Second**, we put forward two improvements of SVM-based relevance feedback: (a) to optimize the transfer of information between the user and the system, we introduce a new active learning selection criterion that minimizes redundancy between the candidate images shown to the user, and (b) as most image classes span a wide range of scales in the description space, we argue that the insensitivity of the SVM to the scale of the data is desirable in this context and we show how to obtain it by using specific kernel functions.

In the following, we position our approach with respect to the existing work and we detail our contributions. We end the section by an outline of the paper.

Hybrid visual and conceptual content representation. If we regard the information provided concerning the target images or the possibilities of interaction between the user and the system, keywords and visual content appear to be rather complementary to each other and it is important to rely on both of them for the retrieval of images.

Keywords associated to an image can be divided in two categories: (i) keywords corresponding to *identifiable items* characterizing the *visual content* of the scene and (ii) keywords concerning the *context* and the *interpretation* of the scene. In some cases, it may be possible to automatically obtain keywords in the first category by image analysis, object detection and classification techniques. An explicit use of such keywords during retrieval, to complement descriptors of the visual appearance, will nevertheless be helpful even in cases where the results of image analysis are unreliable. Keywords in the last category are very unlikely to be automatically extracted from the

images, unless very strong correlations exist in a specific image database between visual appearance and such keywords. However, the contribution of such keywords to the retrieval of relevant images is undeniable.

Since in a database some images might have no text annotation, or their annotations might be considered incomplete, a significant amount of work in recent years focussed on the (semi)automatic annotation of images with keywords and on the extension of existing text annotations to images that do not have keywords associated with them (see for example [1], [14], [29], [23], [41], [43]). In this paper we do not deal with this subject, but with the complementary problem of image content description and retrieval based on the joint use of existing text annotations and visual features.

Indeed, indexing and retrieval approaches relying on keywords and visual features together are of great interest for the semantic gap reduction and were heavily investigated in the recent years, even though significant methodological advances are rare. The techniques developed in the CBIR community for query by example and relevance feedback can be directly applied to the joint use of visual features and annotations if the keywords annotating an image are represented by a vector of fixed dimension (as is usually the case for visual features), which explains why this solution was explored in the literature [24], [29], [37], [42], [45].

Various methods were proposed to obtain a feature vector representation based on the keywords annotating an image. A direct solution is to associate one dimension to every keyword that appears in the annotation of some image in the database [29], [42]. A "soft" representation can also be employed [45], [23]: the value of a feature is seen as a "degree of association" between an image and the keyword. Since the number of different keywords usually increases with the size of the database, this solution does not scale well. It also has difficulties in taking into account synonymy and homonymy (nevertheless, [45] suggests to use an ontology for initializing the similarities between keywords), as well as similarities between the concepts corresponding to different keywords. These problems are partly solved if latent semantic indexing (LSI) is performed on these sparse representations and the resulting low-dimensional vectors are used instead [24]. LSI can also be applied to the joint visual and keyword-based feature vectors, in order to find a hybrid reduced representation [44] that links sets of keywords and images. Unfortunately, to identify meaningful relations between keywords, LSI needs high amounts of data. This requirement can only be met when a relatively large quantity of text—rather than just a few keywords—is associated to every image.

In this paper we propose a hybrid visual and conceptual image representation and retrieval framework for databases where all the images are annotated by keywords. We introduce a new conceptual feature vector based exclusively on the keywords annotating each image. Using an

external lexical database, we derive a set of "key" semantic concepts linked with the keywords employed for annotating the images. For each image in the database we project its keywords on the selected key concepts, obtaining an interpretable vector representation. This feature vector takes into account the conceptual (prior, corpus-independent) relations between keywords and can be used together with the visual feature vector for enhancing the results of a query by visual example or for improving relevance feedback.

Relevance feedback. Relevance feedback (RF) is often used in image retrieval as a tool to refine queries or to define complex, user-dependent classes not easily described in terms of visual features [46]. The RF method embodied in a search engine should operate in real time and should maximize the ratio between the quality (or relevance) of the results and the amount of interaction between the user and the system. An RF method is usually defined by two components, a learner and a selector. At every feedback round, the learner uses the images marked as "relevant" or "irrelevant" by the user to re-estimate the target of the user. Given the current estimation of the target, the selector chooses the images for which the user is asked to provide feedback during the next round.

The task of the learner is very difficult in the context of RF [7], [46], [47] since training examples are scarce (their number is usually lower than the dimension of the description space), the training set is heavily imbalanced (there are often many more "irrelevant" examples than "relevant" ones) and both training and evaluation must be performed in real time. Much recent work is based on support vector machines (SVM) [33] because they avoid too restrictive assumptions regarding the data (e.g. that classes should have elliptic shape), are very flexible (can be tuned by kernel engineering) and allow fast learning and evaluation for medium-sized databases.

In much of the work on RF, the images for which the user is asked to provide feedback at the next round were simply those that were currently considered by the learner as potentially the most relevant; also, in some cases these images are randomly selected. An important step ahead was the introduction in [39] of an *active learning* framework for SVM-based relevance feedback applied to text document retrieval and extended to visual corpora in [38]. We put forward here a new selection criterion that reduces the redundancy between the images for which the user is asked to provide feedback: our criterion encourages the selection of images that are far from each other in the space of low-level visual descriptors and thus allows for a better exploration of the current frontier.

The image classes found in generalist databases (that is, the potential targets of the users of an RF system) can have various shapes and span very different scales in the space of low-level visual descriptors. Learners that are strongly dependent on a scale parameter will then be unable

to find for this parameter a value that is adequate for all classes in the database; it is not possible to tune the scale parameter online, during the few feedback rounds, to suit the query class. Such sensitiveness to the scale of the data does occur for SVMs when some kernels (such as the RBF one) are employed. We argue that invariance (or at least low sensitivity) to scale is then an important desirable feature of the learner in an RF context and we propose to use specific kernel functions which allow the SVM to achieve this.

Outline. In Sec. 2 we introduce our new concept-based content descriptor and in Sec. 3 we present our SVM-based relevance feedback mechanism. We put special emphasis on the selection strategy associated with the RF learner and on the choice of kernels that reduce the sensitivity of the SVM to the scale of the target classes in the description space. In Sec. 4 we present experimental results obtained on a real-world database from the Alinari Picture Library. We discuss some issues deserving further investigation in Sec. 5 and we conclude by a summary of the main achievements of our work.

2 Description of the images

2.1 Visual content descriptors

To account for the visual content of the images, we employ the global color, texture and shape signatures presented in [3]. Our weighted color histograms rely on the use of the Laplacian and of the local probability as pixel weighting functions. These functions bring additional information into the color histograms (such as local shape or texture), which is important for building compact and reliable image signatures. The resulting integrated signatures generally perform better than a combination of classical, single-aspect feature vectors.

The shape content is described by a histogram based on the Hough transform, giving the global behavior along straight lines in different directions. Texture feature vectors are based on the Fourier transform and provide a distribution of the spectral power density along the frequency axes. This signature performs well on textured images and, used in conjunction with other image signatures, can significantly improve the overall behavior. The resulting joint feature vector has more than 600 dimensions. With such a high number of dimensions, relevance feedback may become impractical even for medium-size databases. We use linear principal component analysis to reduce the dimension of the feature vectors about 5 times, with less than 3% loss in the precision/recall diagrams in the query by example framework [16].

2.2 New concept-based content descriptor

We put forward here a new conceptual feature vector for semantic indexing, using the set of keywords annotating an image. This conceptual feature vector provides complementary information both to the relevance feedback mechanism and to the evaluation of the similarity between images in a query by example framework. The dimensions of these vectors are interpreted as similarities to a set of "key" concepts. Having a fixed dimension, our conceptual feature vector can be directly employed by an existing CBIR system.

A simple vector representation for a set of keywords would be to associate a coordinate in the feature vector to every keyword present in the annotation. Not only this solution lacks scalability, but the result of a simple distance computation between such vectors would only depend on the number of keywords shared by the two images and not on the conceptual similarities between *different* keywords.

The main idea behind semantic indexing is to use word senses rather than the words themselves for indexing documents. This may help improving both precision and recall measures by handling different semantic relations between concepts. Related experiments reported in literature show that significant improvements can be obtained in text retrieval [20], [30]. Although it seems desirable for document indexing to take a maximum of semantic information into account, the growth of the number of indexing terms not only increases the processing time, but can also alter precision because of "curse of dimensionality" effects. The problem is not new and various techniques aiming at reducing the size of the indexing set already exist: filtering by part of speech tags, frequencies, or through statistical techniques such as probabilistic latent semantic indexing [22]. The keywords appearing in image databases already correspond to meaningful words. Also, weighting schemes and the search for correlations between keywords are hindered by the fact that most images are annotated by very few keywords, so information collected by statistical means is unreliable. Moreover, LSI-based dimension reduction does not provide interpretable representations (the new coordinates do not correspond to specific concepts).

To obtain a *scalable* solution for representing sets of keywords as *interpretable* feature vectors, we propose to select a limited set of *key concepts* and to associate to every such concept a dimension in the feature vector. We rely on an ontology, defining semantic relations between concepts, to find good candidates for these key concepts and to define the feature vectors for sets of keywords. We start by motivating our choice of a general ontology (WordNet), and then we continue by describing our method for computing the conceptual feature vectors.

WordNet is arguably the most popular and widely used semantic resource in the computational

linguistics community today. It is a database of nouns, verbs, adjectives and adverbs organized into discrete senses (synsets), each representing one underlying lexical concept. The concepts are linked by several semantic relations of various types, such as synonymy, antonymy, hypernymy, hyponymy, holonymy, meronymy, etc. One of the reasons for its success and wide adoption is its ease of use: as a semantic network with words at the nodes, it can be readily applied to any textual input for query expansion, or determining semantic similarity. Further details regarding WordNet can be found in [15].

At this time, some widely available alternatives to WordNet as a general-purpose semantic knowledge database are Cyc and ConceptNet. The Cyc project attempts to formalise commonsense knowledge into a logical framework [26] with assertions largely handcrafted by knowledge engineers Using Cyc to reason about text involves mapping the text into a proprietary logical representation, relying on its own language CycL. However, this mapping process is quite complex because all of the inherent ambiguity in natural language must be resolved to produce the unambiguous logical formulation required by CycL. ConceptNet is a semantic network of commonsense knowledge composed of semi-structured text fragments, linked by an ontology using semantic relations [28]. Rather than manually handcrafting commonsense knowledge, ConceptNet is generated automatically from the English sentences of the Open Mind Common Sense corpus: their website has been launched in year 2000 as an open public collaborative project [35]. While appealing, ConceptNet is still a recent initiative, far from the maturity of WordNet and Cyc, and yet not as widely employed.

Following the above reasons, we use WordNet to build and test our new conceptual descriptor. Nevertheless, our method is in no way specific to WordNet and can be employed with the ontology that is most appropriate for the field of application. For the purpose of our evaluations, recent versions of WordNet provide a fair set of semantic relations such as "IS A", "HAS A" and "IS PART OF", a set of semantic similarity measures based on them as well as the possibility to build the hypernym[1] graph associated with a set of concepts. While WordNet does not yet offer general lateral connections between any two concepts (e.g. "doctor" and "hospital"), we do obtain very good results that could certainly be improved by the use of a more complete knowledge database.

The **key concepts** we need for building the conceptual feature vectors should allow us to evaluate the conceptual similarity between keywords w that are mapped to different concepts $c(w)$ in the ontology. We must then rely on the hypernymy/hyponymy subgraph linking the concepts associated to all the keywords in the database to the most generic concepts (such as "entity").

[1]The concept Y is a hypernym of the concept X if Y carries the meaning of X, i.e. X can be replaced by Y without a change in meaning, but the inverse is not necessarily true.

For every concept corresponding to a keyword annotating an image, we find all the paths in the ontology that lead to the most generic concepts. These paths, obtained for all the keywords in the database, define the hypernyms subgraph we are interested in. A small set (compared to the number of different keywords) of key concepts is then empirically selected. When choosing the key concepts, a certain balance must be obtained with respect to their semantic broadness. When choosing concepts that are too close to the keywords, the set of key concepts will not reduce enough, preserving some distinctions between words with close senses. Choosing concepts that are too generic will reduce more the set of key concepts, but the payoff is a lower power of discrimination. Good key concept candidates are super-concepts of several keywords and are relatively close to these; also, the key concepts must be balanced among all the branches (see Fig. 1 for an example). Further discussion and some ideas about how to automatically choose the set of key concepts are presented in Sec. 5.

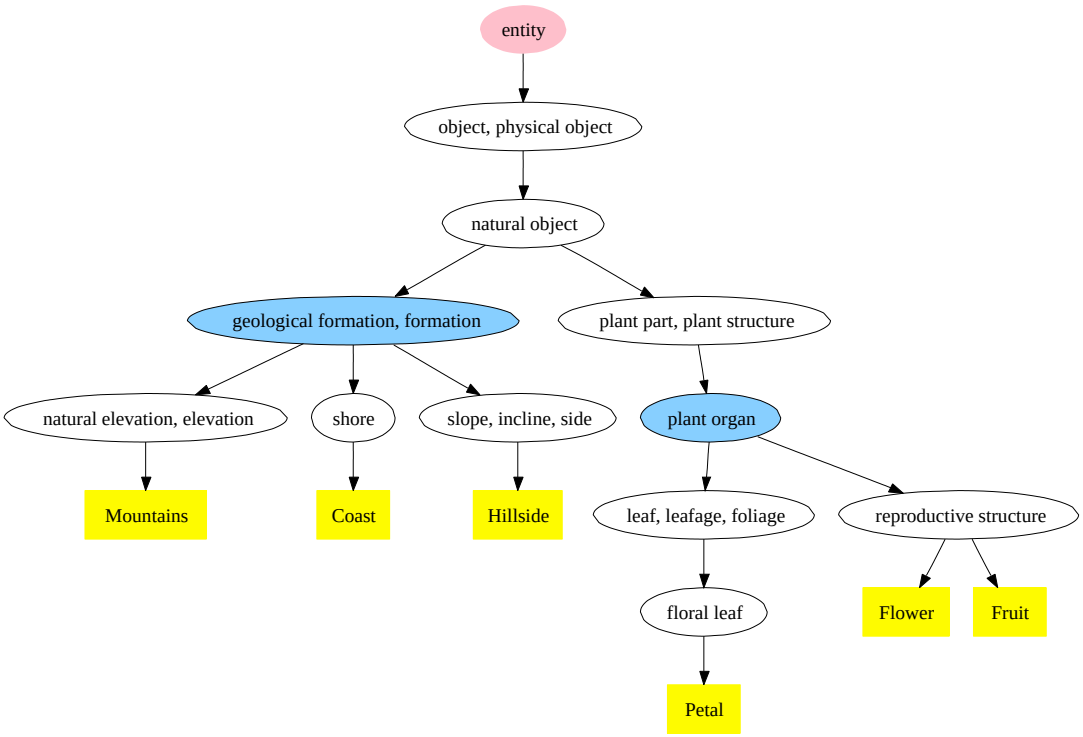


Figure 1: Hypernym graph generated by WordNet for a set of keywords. The nodes "geological formation" and "plant organ" are good key concept candidates: they are parent nodes for several keywords and are relatively close to these. They are not too generic and avoid keeping the distinction between words with close senses.

After selecting the key concepts, we compute for every image a **conceptual feature vector** representing the projection of its keywords. We first study representations for single keywords, then we turn to sets of keywords.

In our feature vector, one dimension is dedicated to every key concept. Suppose that $\{C_i|1 \leq$

$i \leq n\}$ are the n selected key concepts. Let us consider a keyword w mapped to a concept $c(w)$ and denote by $\mathbf{v}(c(w))$ the feature vector representing this keyword alone. A simple solution is to define the components of the feature vector according to

$$v_i(c(w)) = \begin{cases} 1, & \text{if } C_i \text{ is a super-concept of } c(w) \\ 0, & \text{otherwise} \end{cases}$$

This method for computing feature vectors is denoted in the following by WNS-BINARY (the suffix WNS stands for "WordNet Signature"). In this case, the keywords mapped to concepts that are different but have the same key super-concepts will have the same feature vectors. A refined solution should include in $\mathrm{v}(c(w))$ the *similarity* between $c(w)$ and its key super-concepts. Such a similarity can be interpreted as the relevance of a key concept for describing an image annotated with the keyword. We thus have to evaluate similarities between concepts.

There are several **measures of conceptual similarity**, relying on WordNet, that can be used for the definition of our concept-based feature vector. The measures put forward in [25] and [40] rely on knowledge-rich sources (ontologies) alone, while those in [5], [27], [31] combine them with knowledge-poor sources (corpus statistics). Leacock and Chodorow [25] rely on the length of the shortest path following IS-A relations, $\mathrm{len}(c_1, c_2)$, between two concepts c_1 and c_2, to measure their semantic similarity. The length of the path is scaled by the overall depth D of the concept taxonomy: $\mathrm{sim}_{LC}(c_1, c_2) = -\log(\mathrm{len}(c_1, c_2)/2D)$. Wu and Palmer [40] evaluate the similarity according to how close the two concepts are in the concept hierarchy, $\mathrm{sim}(c_1, c_2) = 2N_3/(N_1 + N_2 + 2N_3)$, where if we denote by c_3 the nearest common super-concept (or lowest super-ordinate) of c_1 and c_2, N_1 is the number of nodes in the path from c_1 to c_3, N_2 from c_2 to c_3 and N_3 from c_3 to the root node. For Resnik [31], the similarity between two concepts depends on the extent to which they share information: similarity is defined as the information content of their lowest super-ordinate $\mathrm{lso}(c_1, c_2)$ according to $\mathrm{sim}_R(c_1, c_2) = -\log p(\mathrm{lso}(c_1, c_2))$, where $p(c)$ is the probability of encountering an instance of a concept c in some specific corpus. The proposal in Lin [27] is based on an information-theoretic similarity measure for arbitrary objects. With the notations above, $\mathrm{sim}_L(c_1, c_2) = 2\log(p(\mathrm{lso}(c_1, c_2)))/[\log(p(c_1)) + \log(p(c_2))]$.

Using one of these measures, indicated by the short names LCH (Leacock-Chodorow), WUP (Wu-Palmer), RES (Resnik) and LIN (Lin), we defined two different types of feature vectors $\mathbf{v}(c(w))$ for representing the keyword w mapped to a concept $c(w)$. In the first one, the components of the

feature vector are obtained by projecting $c(w)$ only on its parent key concepts:

$$v_i(c(w)) = \begin{cases} \text{sim}(c(w), C_i), & \text{if } C_i \text{ is a super-concept of } c(w) \\ 0, & \text{otherwise} \end{cases}$$

In a second representation, we do not limit the evaluation of the similarity to the super-concepts of $c(w)$, and we set $v_i(c(w)) = \text{sim}(c(w), C_i)$. Through the similarity measures, this allows the use different semantic relations, other that hypernymy, available in the ontology. In the following, the use of this method will be indicated by the suffix "-ALL" appended after the short name of the similarity measure employed.

If an image I is annotated with the set of keywords $\mathcal{K}(I)$, we define the components of the feature vector $\mathbf{v}(\mathcal{K}(I))$ representing $\mathcal{K}(I)$ as $v_i(\mathcal{K}(I)) = \max_{w \in \mathcal{K}(I)} v_i(c(w))$. Because of the maximum, for every key concept only the keyword whose meaning is closest to this concept has an impact on the feature vector. This avoids the situation where many keywords unrelated to a concept may add-up their contributions to a significant level.

In Sec. 4 we will present experiments with this new hybrid representation, show examples of retrieval together with performance evaluations and study its impact on relevance feedback.

3 Active SVM-based relevance feedback

In this section, we propose an improved SVM-based relevance feedback framework and we discuss the impact of both the learning machine and the selection criterion on the retrieval results. To optimize the transfer of information between the user and the system we define a new active learning selection criterion that minimizes redundancy between the candidate images shown to the user. In addition, since we are confronted to image classes that span very different scales in the description space, we use specific kernel functions to obtain an insensitivity of the SVM to the spatial scale of the data.

3.1 Active learning with reduction of redundancy

In order to maximize the ratio between the quality (or relevance) of the results and the amount of interaction between the user and the system, the selection of images for which the user is asked to provide feedback at the next round must be carefully studied. Cox et al. [11] introduce an information criterion for the *target search* scenario, where the goal is to find a specific image in the database. The user is required to choose between the two images presented by the engine the

one that is closest to the target image. The selection strategy put forward in this case attempts to identify at every round the most informative binary selections, i.e. those that are expected to maximize the transfer of information between the user and the engine (or remove a maximal amount of uncertainty regarding the target). We consider that this criterion translates into two complementary conditions for the images in the selection: each image must be ambiguous given the current estimation of the target and the redundancy between the different images has to be low. But the entropic criterion employed in [11], [10] does not scale well to the search of images in a larger set (*category* search) and to the selection of more than 2 images at every feedback round.

Based on the definition of active learning (see for example [9]), the selection of examples for training SVMs to perform general classification tasks is studied in [6]. In the early stages of learning, the classification of new examples is likely to be wrong, so the fastest reduction in generalization error can be achieved by selecting the example that is farthest from the current estimation of the frontier. During late stages of learning, the classification of new examples is likely to be right but the margin may be suboptimal, so the fastest reduction in error can be achieved by selecting the example that is closest to the current estimation of the frontier. Following the classical formulation of active learning, the authors only consider the selection of single examples for labeling at every round.

Tong and Koller [39] present several selection criteria for SVM learners applied to content-based text retrieval with relevance feedback. The simplest and cheapest of these criteria consists in selecting the texts whose representations are closest to the hyperplane currently defined by the SVM. We shall call this the selection of the "most ambiguous" (MA) candidate(s). This criterion is justified by the fact that knowledge of the label of such a candidate halves the version-space. In this case, the version space is the set of parameters of the hyperplanes in the feature space that are compatible with the already labeled examples. The proof of this result assumes that the version space is not empty and that, in the feature space associated to the kernel, all the images of vectors in the input space have constant norm. These assumptions will hold with appropriate choices for the kernel and for the regularization bound (C). To minimize the number of learning rounds, the user is asked to label several examples at every round and all these examples are selected according to the MA criterion. We note that the MA criterion in [39], [38] is the same as the one put forward in [6] for the late stages of learning. This clarifies the fact that MA relies on two important further assumptions: first, the prior on the version space is rather uniform; second, the solution found by the SVM is close to the center of gravity of the version space. The second assumption can be relieved by using the more sophisticated criteria put forward in [39] or Bayes Point Machines [21] instead of SVMs, albeit at a higher computational cost.

In [38] the MA selection criterion is applied to CBIR with relevance feedback and shown to produce a faster identification of the target set of images than the selection of random images for labeling. While providing a computationally effective solution to the selection of the most ambiguous images, when used for the selection of more than one candidate image the MA criterion does not remove the redundancies between the candidates.

We suggest here to introduce the following additional *condition of low redundancy*: if x_i and x_j are the input space representations of two candidate images, then we require a low value for $K(x_i, x_j)$, which is the value taken by the kernel for this pair of images. If the kernel K is inducing a Hilbert structure on the feature space, if $\phi(x_i)$, $\phi(x_j)$ are the images of x_i, x_j in this feature space and if all the images of vectors in the input space have constant norm, then this additional condition corresponds to a requirement of (quasi-)orthogonality between $\phi(x_i)$ and $\phi(x_j)$ (since $K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$). We shall call this criterion the selection of the "most ambiguous and orthogonal" (MAO) candidates.

The MAO criterion can obviously be extended to reduce redundancies between the examples selected during subsequent RF rounds. This additional constraint may be important in situations where the number of labeled examples is much lower than the dimension of the input space and the classes are restricted in most directions.

MAO has a simple intuitive explanation for kernels $K(x_i, x_j)$ that decrease with an increase of the distance $d(x_i, x_j)$, which is the case for most common kernels: it encourages the selection of unlabeled examples that are far from each other in input space, allowing a better exploration of the current frontier.

To implement this criterion, we first perform an MA selection of a larger set of unlabeled examples. If S is the set of images not yet included in the current MAO selection and x_i, $i = 1 \dots n$ are the already chosen candidates, then we add as a new example the vector $x_j \in S$ that minimizes the highest of the values taken by $K(x_i, x_j)$:

$$x_j \;=\; \text{argmin}_{x \in S} \, \max_i \, K(x, x_i) \tag{1}$$

This condition can be justified by reference to the version space account suggested in [39]: redundancy is minimized when the hyperplanes associated to the individual examples are orthogonal and are thus complementary to each other in halving the version space.

In a general classification context, a rather similar "diversity" condition for the selected examples was put forward in [4] and evaluated on several benchmark classification problems from the UCI database. However, their criterion is computationally expensive and do not scale well to

large image databases. Since most of the images are situated far away from the SVM frontier, it is not necessary to optimize the low redundancy condition over a large set of images. Choosing candidates from a small pre-selection around the frontier, as we propose here, is much faster and does not have a significant impact on the results.

3.2 Sensitivity of the learner to the scale of the data

During the study of several ground truth databases we found that the size of the various classes often covers an wide range of different scales in the space of low level descriptors (1 to 7 in our tests). We expect yet more significant changes in scale to occur from one database to another, from one user-defined image class to another within a large database or even between different parts of the frontier of some classes. A too strong sensitivity of the learner to the scale of the data could then significantly limit its applicability in an relevance feedback context.

Unfortunately, most kernels advocated in the literature for relevance feedback depend on a scale parameter. We include here the RBF kernel, $K(x_i, x_j) = \exp\big(-\gamma\|x_i - x_j\|^2\big)$, highly sensitive to the scale parameter γ and the Laplace kernel, $K(x_i, x_j) = \exp\big(-\gamma\|x_i - x_j\|\big)$, advocated in [8] for histogram-based image descriptors and used in many research works (see [19] for an example).

The triangular kernel, $K(x_i, x_j) = -\|x_i - x_j\|$, was introduced in [2] as a *conditionally* positive definite kernel, but the convergence of SVMs remains guaranteed with this kernel [32]. In [18] the triangular kernel was shown to have a very interesting property: it makes the frontier found by SVMs invariant to the scale of the data.

Although particularly important for relevance feedback, the scale issue has rarely been mentioned explicitly in the literature. We suggest here that kernels inducing a low sensitivity of the learner to the scale of the data, such as the triangular kernel presented above, should be preferred with relevance feedback. Indeed, since the user-defined classes are not known *a priori*, the scale parameter of a kernel cannot be easily adjusted online. Thus, important variations between classes can be expected for the performance of relevance feedback retrieval if kernels such as the RBF one are employed. The scale-invariance obtained by the use of the triangular kernel becomes then a highly desirable feature and, according to the experimental evidence (see Sec. 4), makes this kernel an excellent choice.

3.3 Synthesis of our approach

We conclude this section by a summary of the improvements introduced in our relevance feedback framework:

- We propose a new selection criterion that reduces the redundancy between the selected samples. Also, since most of the images in the database are far away from the SVM frontier, optimizing the selection criterion over all the database is not necessary and slows down the process, making it unpractical even for medium-sized databases. We propose to use a small pre-selection of images around the current frontier and optimize the MAO selection criterion on this pre-selection. Our criterion is as fast as MA [38] while eliminating the redundancy between the selected samples: the overhead imposed is O(constant), independent of the size of the database, unlike the algorithm proposed in [4].

- We address the important issue of the scale of the data and we propose the use of specific kernel functions, such as the triangular kernel, to provide invariance to the scale of the image classes. MAO using this type of kernels is able to learn complex classes of images defined in terms of high level semantic terms. Our method does not require prior knowledge about the query concept and works for generic image databases, even in the absence of keyword annotations.

- For conditionally positive definite kernels (such as the triangular kernel we are using in our experiments), the account provided in [39] and [38] for the MA selection criterion does not hold. However, since the value of $K(x_i, x_j)$ decreases with an increase of the distance $d(x_i, x_j)$, our justification for the MAO criterion holds.

4 Experimental evaluation

In this section we present experimental evaluations validating our approach. In Sec. 4.1 we introduce the experimental setup and the performance measures we use. In Sec. 4.2 we compare our relevance feedback mechanism with alternative methods and in Sec. 4.3 we evaluate the joint use of the visual descriptors with our new concept-based feature vector.

4.1 Experimental setup

Ground truth database. We start from a large image collection kindly provided by Alinari[2], containing images related to art, archeology, architecture, etc. We first selected as test database the 3585 images that are annotated by at least two keywords. Then, we built by hand a ground truth that is consistent with both the visual aspect and the higher-level semantics of the images. Note that no ground truth class is the union or intersection of sets of images annotated with the

[2] www.alinari.com

same keyword(s). This is to assure that the results are not biased by giving an unfair advantage to the conceptual feature vector (which is build using the keyword annotations). We defined 20 classes in the ground truth, having between 15 and 174 images each. The number of files included in the ground truth is 1073 and the mean overlapping between classes is of about 10%. A certain degree of overlapping between ground truth classes corresponds better to real situations where an image may belong to several different user-defined image classes. While the ground truth is smaller than the test database, we perform all the evaluations on the entire test database of 3585 files. The evaluation procedure only regard the relevance of the images belonging to the ground truth, the other images are considered neutral to the query. However, the images that do not belong to the ground truth may act like noise making the task of the system more difficult, which corresponds better to a real world situation.

The concept-based feature vector. Following the procedure presented in Sec. 2.2, we built the hypernym graph associated with the whole test database and we selected 28 representative key concepts to be used for projecting the sets of keywords that annotate the images. Thus, the concept-based feature vector has 28 dimensions. No keyword was included as a key concept. To represent the visual content of the images, we use the visual feature vector described in Sec. 2.2.

4.2 Evaluation of the relevance feedback mechanism

We evaluate our relevance feedback mechanism, introduced in Sec. 3, on the ground truth image database described above. We compare our new selection criterion (MAO: Most Ambiguous and Orthogonal) with the MA criterion (Most Ambiguous) proposed in [38] and with the standard selection criterion used in image retrieval with relevance feedback: select the unlabeled examples for which the current decision function of the SVM has the highest positive values (denoted here by MP: Most Positives). We also show results comparing several kernels with the triangular kernel, results that illustrate well the problems introduced by a strong sensitivity of the learner to the scale of the data.

At every feedback round the emulated user labels as "relevant" or "irrelevant" all the images in a window of size $ws = 9$. Every image in every ground truth class serves as the initial "relevant" example for a different RF session, while the associated initial $ws - 1$ "irrelevant" examples are randomly selected. The target of each RF session is to find all images in the class where the initial positive example belongs. When we use the MAO selection criterion, it is computed on a window of size $2 \times ws$. For all the kernels we employed the L1 norm because experimentally we found it to provide better results than L2.

We follow each relevance feedback session for 30 iterations (rounds) and we measure the preci-

sion within a window of the size of the class. This window size gives the system a chance to achieve the perfect recall, $R = 1$. Since we perform an exhaustive testing by starting a RF session for each image in every class, at every iteration we compute the mean value of the precision measure over all feedback sessions. This provides an average measure of how well RF performs, iteration by iteration, in its task of finding the target class. As image features, we employ a combination of the visual features and the WNS-LIN-ALL signature (projection of keywords on all key concepts through the Lin semantic similarity measure) introduced in Sec. 2.2.

Sensitivity of the learner to the scale of the data. First, we evaluate the sensitivity of the kernels presented in Sec. 3.2 to the scale of the classes of images included in the ground truth. We use several values for the scale parameter and for each diagram we take the mean value of the precision for the first 30 feedback iterations. This is a measure of how well RF performs with respect to the ground truth for the given scale parameter. In Fig. 2 we present the results obtained for seven values of the scale parameter, $\gamma \in \{0.001, 0.01, 0.1, 1, 10, 100, 1000\}$ vs. the behavior of the triangular kernel (having no scale parameter).

The RBF kernel is the one producing the highest sensitivity to scale for the SVM. Since no scale parameter value is convenient for all classes in the ground truth, the performance of the RBF kernel never approaches that of the triangular kernel. For the Laplace kernel, an increase of γ beyond 1 has a strong negative impact on the results, while a reduction of γ does not have significant consequences. This is explained by the fact that for small γ the Laplace kernel becomes similar to the triangular kernel. The invariance to scale obtained by the use of the triangular kernel proves to be a very useful property for generalist databases, when the target class is complex and best described in semantic terms (see also [17] for experimental evaluations involving several other databases and using only the visual descriptors).

Comparison of the selection strategies. We compare here our selection criterion, MAO, with other alternative popular selection criteria employed in the literature (MA and MP). We also discuss how active learning selection criteria are expected to work with different types of image classes. To match as closely as possible a real context, our ground truth is built from high level image classes, independent of the keyword annotations and visual appearance, all classes having a high range of internal diversity. We built mean precision vs. iteration diagrams, as explained above, for each class in our ground truth. By examining the results on each class, we found two main types of behavior.

The first situation is illustrated in Fig 3. The target class "non human statue" is very abstract and was build by choosing all images depicting statues or parts of statues in the database, excluding human statues. There is a large variety of shapes, colors and textures involved in the visual

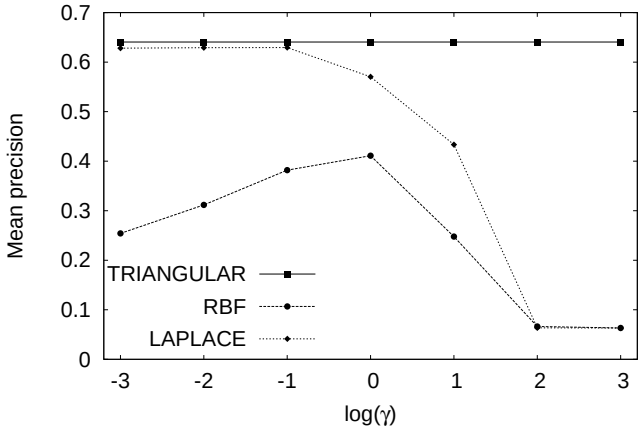


Figure 2: SVM sensitivity to the scale of the data: mean precision vs. scale parameter diagram for different kernels using the MAO selection criterion. Kernels depending on a scale parameter (the RBF and Laplace kernels in the diagram) are highly sensitive to the scale parameter. Since the user defined image classes are not known *a priory* and the scale parameter cannot easily be adjusted online, the scale invariance provided by the triangular kernel becomes a highly desirable feature.

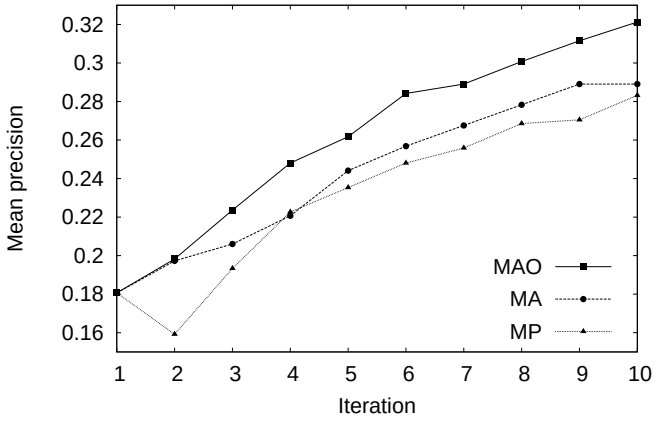


Figure 3: Comparison of selection strategies: the target class ("non human statue") is not well separated in the description space. In this situation, MA improves the results only by a small margin. However, due to eliminating the redundancies between the selected samples, the MAO criterion yields better results.

description of the class, as well as a set of unrelated keywords (names of animals and mythological creatures). As we can see, the MA criterion improves somehow the results compared to MP, but not by a very large margin (around 1%). This is explained by the fact that the class is not not well separated from the rest of the database in the description space and, thus, the class frontier is not well defined. However, our criterion (MAO), by eliminating the redundancy between the selected samples, offer results that are consistently better at all feedback iterations (around 4%).

In the second type of behavior, the target class is well separated in the description space. The frontier is well defined and we expect the active learning criteria (MA, MAO) to find it much faster compared to the standard criterion MP, that focuses on the interior of the class. In Fig. 3

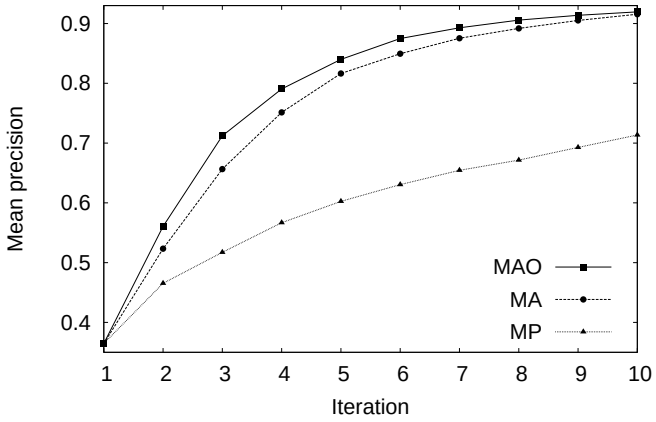


Figure 4: Comparison of selection strategies: the frontier of the target class ("archaeological site") is relatively well defined. In this case, the active learning criteria (MAO and MA) provide significantly better results compared to the standard MP criterion. Moreover, the MAO criterion visibly improves the results, by minimizing the redundancy between the selected samples.

we present results on the class "archaeological site", that illustrates well this situation. After only a few feedback iterations, the active learning criteria improves the results by a large margin, due to exploring the "ambiguous" regions around the estimated frontier in the description space. Moreover, since MAO eliminates the redundancy between the selected samples, it yields visibly better results compared to the MA criterion (approx. 7% after 3 iterations).

We found a similar behaviors for all classes in our ground truth, results that both confirm and extend the evaluations presented in [17] for several ground truth databases, but using only the visual descriptors. To conclude this section, we summarize our findings as follows:

1. Subject to the complexity and the internal structure of the target class, active learning criteria may not always provide better results compared to the standard MP criterion. Experimental evidence suggest that the best situation when active learning improves significantly the results, is for target classes relatively well separated from the rest of the database and having a high internal diversity. In this case, the class frontier is relatively well defined and active learning helps exploring it in an optimized way.

2. Our selection criterion, MAO, systematically surpassed in our experiments the MA and MP criteria, sometimes by a significant margin, as explained above. Since the MAO criterion does not optimize its selection of images over the entire database, its impact on the computation time is low and does not grow with the size of the database. MAO can then be systematically preferred to the MA criterion because it guarantees a reduction of redundancy between the selected samples.

4.3 Evaluation of the combined visual and conceptual descriptor

We present here the evaluation of the joint use of visual features and the concept-based signatures both in a Query By Example (QBE) context and with relevance feedback.

Comparisons in the QBE framework. We test several types of conceptual feature vectors presented in Sec. 2.2 and we build precision-recall diagrams using the ground truth described previously. In the following, the names of the different concept-based descriptors are composed of the base name WNS (standing for "WordNet Signature"), a middle part indicating the semantic similarity measure used to build the descriptors (LCH: Leacock-Chodorow, WUP: Wu-Palmer, RES: Resnik, LIN: Lin, BINARY: binary) and an optional suffix (-ALL) indicating if keyword projections have been done on all key concepts or only on the parent key concepts (see Sec. 2.2).

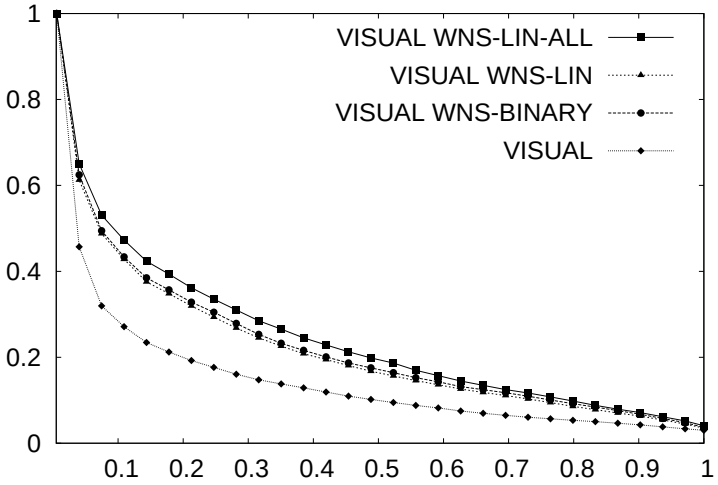


Figure 5: Precision/recall diagrams for QBE with several combinations of VISUAL, WNS-BINARY, WNS-LIN and WNS-LIN-ALL descriptors. Projecting keywords on all the key concepts (WNS-LIN-ALL descriptor) allows the use of semantic relations other than hypernymy, through the similarity functions, which has a positive influence on the results returned by the system.

In Fig. 5 we present precision/recall diagrams for the joint use of visual and WNS-BINARY, WNS-LIN or WNS-LIN-ALL descriptors, and for the visual feature vector alone. The WNS-LIN-ALL signature performs clearly better than WNS-LIN and WNS-BINARY when combined with the visual features, and much better than the visual feature vector alone. We obtained similar diagrams for the Leacock-Chodorow, Wu-Palmer and Resnik similarity measures. These findings were verified throughout the tests we performed in the QBE scenario: (a) using both visual and concept-based feature vectors visibly improves the quality of the results compared to using visual features alone, and (b) projecting the keywords on all the key concepts (WNS-LIN-ALL in the Fig. 5) gives better performance than projecting only on their key super-concepts. However, we could not obtain experimental evidence to favor any of the similarity measures presented in

Sec. 2.2 (see Fig. 5 for an example: the diagrams corresponding to WNS-LIN and WNS-BINARY are roughly identical). For further tests, we choose to use the Lin similarity measure since in [5] it was shown to be the closest to the way human subjects judge similarity.

Comparisons with relevance feedback. We tested our new concept-based feature vector using relevance feedback on the ground truth database introduced in Sec 4.1. The comparisons were performed using the MAO selection criterion, described in Sec. 3, and we employed the triangular kernel for the SVM.

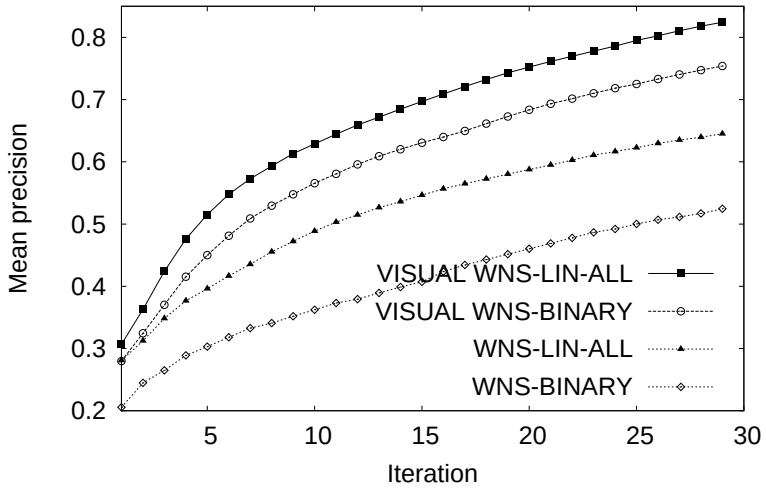


Figure 6: Comparing BINARY and WNS-LIN-ALL feature vectors with relevance feedback: concept-based signatures relying on a semantic similarity measure work much better than the binary projection, both alone and in combination with the visual features.

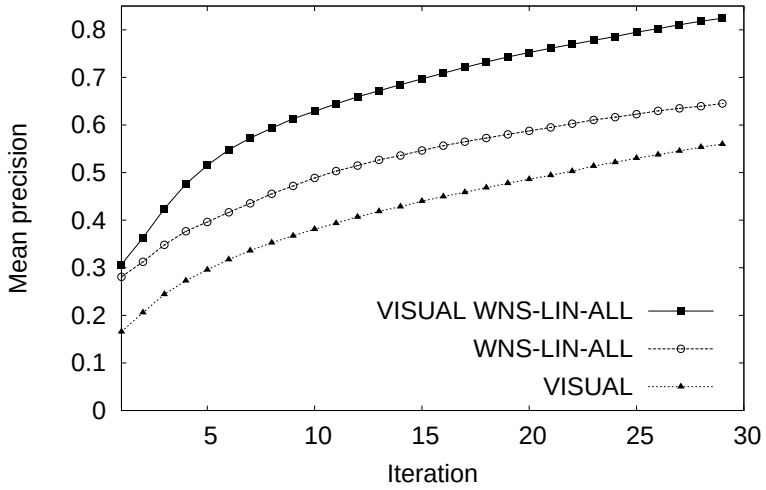
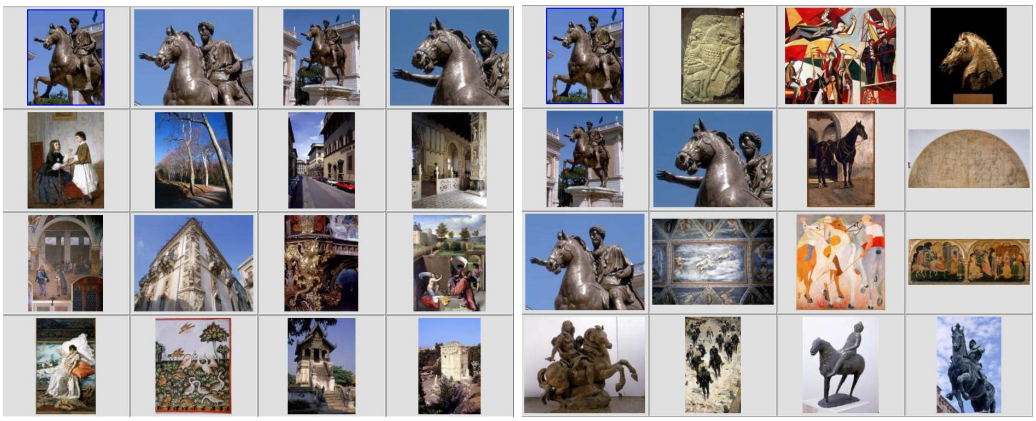


Figure 7: Mean precision vs. iteration diagrams for relevance feedback: the joint use of the visual and concept-based descriptors significantly outperforms their individual use.

Fig. 6 shows that WNS-LIN-ALL significantly outperforms the WNS-BINARY version with

relevance feedback, both when considered alone and when combined with the visual feature vector. Relying on a semantic similarity measure proves to be benefic, since it allows the use of the different semantic relations available in the ontology. Fig. 7 presents mean precision vs. iteration diagrams for the WNS-LIN-ALL signature, employed alone or in combination with the visual feature vector. We see that the joint use of the concept-based signature and the visual feature vector produces a significant improvement of the results, compared to the use of visual or concept-based signatures alone.

The tests performed with relevance feedback complete and reinforce the conclusions of the QBE evaluation: (a) the concept-based signatures relying on semantic similarity measures presented in Sec. 2.2 work better than the binary concept-based signature and (b) projecting the keywords on all the key concepts provides better results than employing only the key super-concepts. Also, relevance feedback performance is much better when employing the joint visual and conceptual feature vectors. This is an indication of the fact that user feedback allows the system to make a better use of the information provided separately by the two types of descriptors, selecting at each iteration what is most useful in the identification of the target.



(a) VISUAL (b) WNS-LIN-ALL

Figure 8: First page of QBE retrieval results with (a) the visual descriptor and (b) the concept-based descriptor.

As an illustration, in Fig. 8 and Fig. 9 we present some screens of results returned by our system in a QBE scenario: the query image is in the top-left corner of the screenshots and the user is searching for open air statues featuring a knight and a horse. In Fig. 8 (a) we see the results when the system is using only the visual features; in this case the system is confused by too many images in the database having similar visual descriptors with the query image (the semantic gap). The results in Fig. 8 (b) correspond to the use of the WNS-LIN-ALL signature alone: while the returned images are conceptually related to the query image, many of them are too abstract

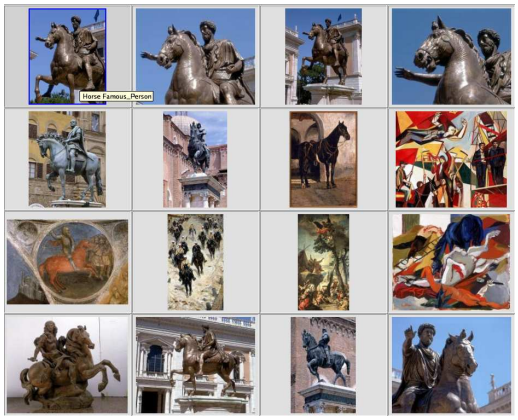


Figure 9: First page of QBE retrieval results with the combined visual and concept-based descriptors.

(paintings, archaeological items) and does not always represent well user's intent. Fig. 9 shows the results obtained when employing both visual and concept-based descriptors. In this case, the system return many images that are not present in the previous situations and which clearly correspond better to the target class.

5 Discussion and perspectives

In this paper we presented an integrated approach to the semantic gap for image databases containing text annotations. Our approach combine a hybrid conceptual and visual content description together with an improved relevance feedback scheme. In this section we discuss several open issues and we suggest some directions to further develop our work.

Our retrieval framework is primarily intended for databases where all the images have text annotations, but it can also be used with no modifications for databases without text annotations. Indeed, since we employ an active learning scheme based on a kernel that reduce the sensitivity of the learner to the scale of the data, our relevance feedback scheme can approach with very good results complex image classes without the need of external annotations [17]. However, using partially annotated databases raises several issues. To easily perform both queries by example and relevance feedback, the descriptors of all the images should belong to the same vector space. A simple answer to this issue consists in adding to the visual descriptor of the annotation-free images a conceptual vector where all the components are set to zero or to the unconditional expected value. A better solution is to use methods (such as those presented in the introduction) for extending existing annotations to those images that are not annotated, then build the conceptual feature vectors as for the annotated images.

To build our conceptual descriptor, the empirical selection of the key concepts is feasible for

rather small databases, but an automatic method is needed for very large databases given the potential complexity of the hypernym graph. This is a difficult task: by choosing too general concepts the precision of the system will degrade, while by selecting too specific concepts the indexing set will not reduce enough, preserving some distinction between words with close senses (lack of recall). The criteria for selecting nodes in the hypernym graph as key concepts should include the number of relevant children (i.e. concepts found in the annotations) and the distances to these children. An interesting criterion, minimum redundancy cut, was put forward in [34] and is based on information theory; the idea is to choose the appropriate level of conceptual indexing by considering minimal cuts in the hypernym graph, i.e. cuts defining a well-balanced coverage of all the relevant nodes (corresponding to the keywords). However, this entropy-based criterion is not well adapted to our context, and we pursue the investigation of alternative solutions.

Regarding the feedback mechanism, it is clear that in the early stages of the learning, due to the small number of training samples, the frontier may be quite unreliable. As can also be seen from [6], using MA (or MAO) at this stage may prove suboptimal. A promising strategy would be to use MP at the beginning of an RF session and then, when the frontier becomes more reliable, continue with MAO. To switch from MP to MAO, one possibility is to check the evolution of the SVM learner: when the support vectors stay mostly the same from one iteration to the next, the learner is exploring only the inside region delimited by the frontier of the SVM, so it is time to switch to MAO.

Little was said regarding the exact nature of the complexity of image classes and the impact of this complexity on the performance of RF. In [19] the authors propose to adapt the active learning process during RF to the complexity of the target image classes, described by three measures: scarcity, isolation and diversity. Since the complexity is not known *a priori* for user-defined classes, the authors propose to estimate it with the help of existing text annotations. However, we could notice high variations in complexity for different classes in a single image database. The results presented here, together with those shown in [12] for classes having several distinct modes, suggest that such an adaptation may not be necessary if the triangular kernel is employed, since it provides robustness to variations in the complexity of the target class.

6 Conclusion

Although image retrieval using low-level visual features can perform well in many situations, its application to generalist image databases is often limited by the semantic gap. Alternatively, image annotations (keywords) are more directly related to the high-level semantics of the images

and may not reflect well visual similarities. Keywords and visual features provide complementary information, coming from different sources, so the ability of using them together is an advantage in many applications.

In this paper we introduced a new conceptual feature vector that provides reliable similarities between sets of keywords by making use of an external ontology to induce a semantic generalization of the concepts corresponding to the keywords. This conceptual descriptor can be easily combined with visual descriptors for answering queries by example and for performing retrieval with relevance feedback.

We also put forward an improved relevant feedback mechanism employing a new selection criterion based on active learning for SVMs with a reduction of the redundancy between samples. Since user-defined image classes show large differences in scale in the description space, we propose to employ specific kernels that reduce the sensitivity of the SVM to the scale of the data. As shown by the experimental evaluation performed on a ground truth built from a real generalist image database, the joint use of the conceptual content representation together with our new relevance feedback framework contribute to a significant improvement of the retrieval results.

Acknowledgements

We are very grateful to Fratelli Alinari (`http://www.alinari.com`) and especially to Andrea de Polo for providing us the annotated image database.

References

[1] W. H. Adams, G. Iyengar, C.-Y. Lin, M. R. Naphade, C. Neti, H. J. Nock, and J. R. Smith. Semantic indexing of multimedia content using visual, audio and text cues. *EURASIP Journal on Applied Signal Processing*, 3(2):170–185, 2003.

[2] C. Berg, J. P. R. Christensen, and P. Ressel. *Harmonic Analysis on Semigroups*. Springer-Verlag, 1984.

[3] N. Boujemaa, J. Fauqueur, M. Ferecatu, F. Fleuret, V. Gouet, B. L. Saux, and H. Sahbi. Ikona: Interactive generic and specific image retrieval. In *Proceedings of the International workshop on Multimedia Content-Based Indexing and Retrieval (MMCBIR'2001)*, 2001.

[4] K. Brinker. Incorporating diversity in active learning with support vector machines. In *Proceedings of ICML-04, International Conference on Machine Learning*, pages 59–66, August 2003.

[5] A. Budanitsky and G. Hirst. Semantic distance in wordnet: An experimental, application-oriented evaluation of five measures. In *Proceedings of the Workshop on WordNet and Other Lexical Resources NAACL 2001*, 2001.

[6] C. Campbell, N. Cristianini, and A. Smola. Query learning with large margin classifiers. In *Proceedings of ICML-00, 17th International Conference on Machine Learning*, pages 111–118. Morgan Kaufmann, 2000.

[7] E. Y. Chang, B. Li, G. Wu, and K. Goh. Statistical learning for effective visual image retrieval. In *Proceedings of the IEEE International Conference on Image Processing (ICIP'03)*, pages 609–612, September 2003.

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

[9] D. A. Cohn, Z. Ghahramani, and M. I. Jordan. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4:129–145, 1996.

[10] I. J. Cox, M. L. Miller, T. P. Minka, T. Papathomas, and P. N. Yianilos. The Bayesian image retrieval system, PicHunter: theory, implementation and psychophysical experiments. *IEEE Transactions on Image Processing*, 9(1):20–37, January 2000.

[11] I. J. Cox, M. L. Miller, S. M. Omohundro, and P. N. Yianilos. An optimized interaction strategy for Bayesian relevance feedback. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 553–558. IEEE Computer Society, 1998.

[12] M. Crucianu, J.-P. Tarel, and M. Ferecatu. A comparison of user strategies in image retrieval with relevance feedback. In *Proc. of the 7th International Workshop on Audio-Visual Content and Information Visualization in Digital Libraries (AVIVDiLib'05)*, pages 121–130, May 2005.

[13] A. del Bimbo. *Visual Information Retrieval*. Morgan Kaufmann, 1999.

[14] P. Duygulu, K. Barnard, J. F. G. de Freitas, and D. A. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Proceedings of the 7th European Conference on Computer Vision-Part IV*, pages 97–112. Springer-Verlag, 2002.

[15] C. Fellbaum and G. Miller, editors. *WordNet: An Electronic Lexical Database*. The MIT Press, 1998.

[16] M. Ferecatu. *Image retrieval with active relevance feedback using both visual and keyword-based descriptors*. PhD thesis, INRIA—Université de Versailles Saint Quentin en Yvelines, France, 2005.

[17] M. Ferecatu, M. Crucianu, and N. Boujemaa. Retrieval of difficult image classes using svm-based relevance feedback. In *Proceedings of the 6th ACM SIGMM International Workshop on Multimedia Information Retrieval*, pages 23 – 30, October 2004.

[18] F. Fleuret and H. 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.

[19] K. Goh, E. Chang, and W. Lai. Multimodal concept-dependent active learning for image retrieval. In *ACM International Conference on Multimedia 2004*, pages 564–571, 2004.

[20] J. Gonzalo, F. Verdejo, I. Chugur, and J. Cigarran. Indexing with wordnet synsets can improve text retrieval. In *Proc. of the COLING/ACL 1998 Workshop on Usage of WordNet for Natural Language Processing*, pages 38–44, 1998.

[21] R. Herbrich, T. Graepel, and C. Campbell. Bayes point machines. *Journal of Machine Learning Research*, 1:245–279, 2001.

[22] T. Hofmann. Probabilistic latent semantic indexing. In *Proc. of the 22th International Conference on Research and Development in Information Retrieval (SIGIR'99)*, pages 50–57, 1999.

[23] M. Kherfi, D. Brahmi, and D. Ziou. Combining visual features with semantics for a more effective image retrieval. In *Proc. of the 17th International Conference on Pattern Recognition*, 2004.

[24] M. La Cascia, S. Sethi, and S. 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.

[25] C. Leacock, M. Chodorow, and G. A. Miller. Using corpus statistics and WordNet relations for sense identification. *Computational Linguistics*, 24(1):147–165, 1998.

[26] D. Lenat. Cyc: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38, 1995.

[27] D. Lin. An information-theoretic definition of similarity. In *Proc. 15th International Conf. on Machine Learning*, pages 296–304, 1998.

[28] H. Liu and P. Singh. Conceptnet: a practical commonsense reasoning tool-kit. *BT Technology Journal*, 22(4):211–226, 2004.

[29] Y. Lu, C. Hu, X. Zhu, H.-J. Zhang, and Q. Yang. A unified framework for semantics and feature based relevance feedback in image retrieval systems. In *Proceedings of the 8th ACM International Conference on Multimedia*, pages 31–37. ACM Press, 2000.

[30] R. Mihalcea and D. Moldovan. Semantic indexing using wordnet senses. In *Proc. of ACL Workshop on IR and NLP*, 2000.

[31] P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. In C. S. Mellish, editor, *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, pages 448–453, San Mateo, Aug. 20–25 1995. Morgan Kaufmann.

[32] B. Schölkopf. The kernel trick for distances. In *Advances in Neural Information Processing Systems*, volume 12, pages 301–307. MIT Press, 2000.

[33] B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, 2002.

[34] F. Seydoux and J.-C. Chappelier. Semantic indexing using minimum redundancy cut in ontologies. In *Proc. of International Conference on Recent Advances in Natural Language Processing (RANLP 2005)*, pages 486–492, September 2005.

[35] P. Singh, T. Lin, E. Mueller, G. Lim, T. Perkins, and W. Zhu. Open mind commonsense: knowledge acquisition from the general public. In *Proceedings of the First International Conference on Ontologies, Databases, and Applications of Semantics for Large Scale Information Systems*, 2002.

[36] 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.

[37] J. R. Smith, S. Basu, C.-Y. Lin, M. R. Naphade, and B. Tseng. Integrating features, models and semantics for content-based retrieval. In *Proceedings of the international workshop on MultiMedia Content-Based Indexing and Retrieval (MMCBIR'01)*, pages 95–98, September 2001.

[38] S. Tong and E. Chang. Support vector machine active learning for image retrieval. In *Proceedings of the 9th ACM international conference on Multimedia*, pages 107–118. ACM Press, 2001.

[39] S. Tong and D. Koller. Support vector machine active learning with applications to text classification. In *Proceedings of ICML-00, 17th International Conference on Machine Learning*, pages 999–1006. Morgan Kaufmann, 2000.

[40] Z. Wu and M. Palmer. Verb semantics and lexical selection. In *32nd. Annual Meeting of the Association for Computational Linguistics*, pages 133–138, New Mexico State University, Las Cruces, New Mexico, 1994.

[41] C. Zhang and T. Chen. An active learning framework for content-based information retrieval. *IEEE Transactions on Multimedia*, 4(2):260–268, 2002.

[42] H.-J. Zhang and Z. Su. Improving CBIR by semantic propagation and cross-mode query expansion. In *Proceedings of the international workshop on MultiMedia Content-Based Indexing and Retrieval (MMCBIR'01)*, pages 83–86, September 2001.

[43] R. Zhang, Z. M. Zhang, M. Li, W.-Y. Ma, and H.-J. Zhang. A probabilistic semantic model for image annotation and multi-modal image retrieval. In *Proc. of the 2005 IEEE International Conference on Computer Vision (ICCV'05)*, 2005.

[44] R. Zhao and W. I. Grosky. Narrowing the semantic gap—improved text based web document retrieval using visual features. *IEEE Transactions on Multimedia*, 4(2):189–200, 2002.

[45] X. S. Zhou and T. S. Huang. Unifying keywords and visual contents in image retrieval. *IEEE Multimedia*, 9(2):23–33, 2002.

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

[47] Z.-H. Zhou, K.-J. Chen, and Y. Jiang. Exploiting unlabeled data in content-based image retrieval. In *Proc. of the 15th European Conference on Machine Learning (ECML'04)*, pages 525–536, 2004.