Hybrid Visual and Conceptual Image Representation in an Active Relevance Feedback Context

Marin Ferecatu
INRIA Rocquencourt
BP 105 Rocquencourt
78153 Le Chesnay
Cedex, France
Marin.Ferecatu@inria.fr

Nozha Boujemaa
INRIA Rocquencourt
BP 105 Rocquencourt
78153 Le Chesnay
Cedex, France
Nozha.Boujemaa@inria.fr

Michel Crucianu
INRIA Rocquencourt
BP 105 Rocquencourt
78153 Le Chesnay
Cedex, France
Michel Crucianu@inria.fr

ABSTRACT

Many of the available image databases have keyword annotations associated with the images. In spite of the availability of good quality low-level visual features that reflect well the physical content, image retrieval based on visual features alone is subject to semantic gap. Text annotations are related to image context or semantic interpretation of the visual content and are not necessarely directly linked to the visual appearance of the images. Keywords and visual features thus provide complementary information. Using both sources of information is an advantage in many applications and recent work in this area reflects this interest. In this paper, we address the challenge of semantic gap reduction using a hybrid visual and conceptual representation of the content within an active relevance feedback context. We introduce a new feature vector, based on the keyword annotations available for the images, which makes use of conceptual information extracted from an external lexical database, information represented by a set of “core concepts”. Our experiments show that the use of the proposed hybrid conceptual and visual feature vector dramatically improves the quality of the relevance feedback results.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Relevance feedback*

General Terms

Algorithms, Experimentation, Performance.

Keywords

Image retrieval, image descriptors, keyword annotations, relevance feedback.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
MIR'05, November 10–11, 2005, Singapore.
Copyright 2005 ACM 1-59593-244-5/05/0011 ...$5.00.

1. INTRODUCTION

The amount of available multimedia documents has steadily increased lately and with it the need for efficient organization and retrieval of this information when needed. Simple arrangements of items in the database and immediate lookup is no longer sufficient with users more interested by the content than by the description tags found in most archives. These growing needs have boosted research activities in the field of content-based image retrieval (CBIR). Hence, besides these human-based metadata (text) that usually bring semantic information, machine-based meta-data related to the physical content and its low-level features become available for information retrieval [22], [7], [15].

In query by visual example (QBVE), the retrieval results express an overall global visual similarity, thus an approximate similarity. For many visual queries, this leads to differences between user intention/target and the retrieved results. The concept of semantic gap is widely used in content based image retrieval to express the discrepancy that exists between the low level visual features extracted from the images and descriptions that are meaningful to the user [12]. If we consider either the information provided regarding the target concept or the possibilities of interaction between the user and the system, keywords and visual content appear to be complementary to each other and is valuable to rely on both of them for retrieval. Combined image and text indexing and retrieval approaches are of great interest for the semantic gap reduction and are heavily investigated lately.

Part of the work attempting to establish a relation between keywords and visual content consists in the modeling of the visual appearance of images or of image regions corresponding to given concepts. In [8], the authors are searching for a correspondence between image *regions* and keywords that were only provided for *entire* images but refer to regions; the method is based on the development (using expectation maximization) of a joint statistical model of the occurrence of keywords and low-level visual descriptions. Hierarchical aspect models and latent Dirichlet allocation are evaluated in [2], where the authors also study the extension of annotations to other entire images.

Supervised learning is used in [1] (see also [23]) for obtaining models (Markov models or support vector machines) of the “visual content” of “atomic concepts” that can be objects, scenes or events and are associated to keywords. In [17], descriptions of image regions are directly associated to user-provided rough visual descriptions—in terms of color, position, size, shape—of concepts in an ontology.

In [13], vectorial representations are produced for the texts associated to images and *latent semantic indexing* is performed. Every image is then described both by a vector of visual features and by the latent semantic index (vector) of the text associated to the image. The resulting feature vector is based on statistical information computed from the available corpus, so it strongly depends on the quantity and relevance of the available textual data. The presence of joint representations (including both visual and textual features) makes *combined* search possible, often using some form of relevance feedback (RF) as in [26].

In [27], the authors use relevance feedback with the visual features to create a set of synonymy relations linking the keywords used in image annotations. This *a posteriori* approach relies on the consistency of the image classes found by RF, and also on how well the visual features reflect the relations between keywords. Moreover, large images databases are usually annotated with thousands of different keywords: standard keyword vector descriptors are thus very high dimensional, fact that limits their use with RF.

Our approach is complementary to the work described above: as textual annotations available for an image are not sufficient to provide a reasonably accurate statistical description of the relation between words, we use a conceptual generalization approach, based on an external ontology, to derive *a priori*, independently of RF, a low dimensional vector of "core concepts" that is capable of representing a large number of keywords and takes into account their their conceptual relations. Because of its small footprint, our conceptual descriptor can be effectively used with relevance feedback for large generalist image databases. We deal with databases where every image is annotated by some keywords and we introduce a new concept-based feature vector relying on core concepts extracted from keyword annotations.

We use WordNet[9] as an external ontology and we derive a set of core semantic concepts linked with the keywords used for annotating all the images. For each image in the database we project the keywords in its annotation on the selected core concepts obtaining a vector representation. This feature vector can be used as any other image descriptor, for enhancing the results of a query by example or for improving relevance feedback. In Sec. 2.1 we present the visual features we employ and in Sec. 2.2 we introduce our new concept-based feature vector. In section 3 we describe our relevance feedback framework and in Section 4 we present experimental results obtained on a real-world database from the Alinari Picture Library. We conclude the paper by a summary of the main achievements of our approach.

2. DESCRIPTION OF THE IMAGES

We start by a brief presentation of the visual descriptors we employ and then we introduce our new keyword-based conceptual descriptor.

2.1 Visual content descriptors

To describe the visual content of the images we employ the weighted color histograms described in [4], using the Laplacian and local probability as pixel weighting functions. Weighting functions bring additional information into the histograms (e.g. local shape, texture), which is important for building compact and reliable image descriptors. The resulting integrated signatures generally perform better than a combination of classical, single-aspect descriptors. To describe the shape content of an image we use a histogram based on the Hough transform, which gives the global behavior along straight lines in different directions. Texture feature vectors are based on the Fourier transform, obtaining a distribution of the spectral power density along the frequency axes. This descriptor performs well on texture images and, used in conjunction with other image descriptors, can significantly improve the overall behavior. We use linear principal component analysis to reduce the number of dimensions more than 5 times, with a less than 2% loss of performance on the precision/recall diagrams for several ground truth databases.

2.2 New concept-based descriptor

We put forward here a new conceptual feature vector based on the set of keywords that annotate an image. This new feature vector provides complementary information both to the relevance feedback (RF) mechanism and to the evaluation of the similarity between images in a query by example (QBE) framework. With such a feature vector representation, the conceptual information brought in by the annotations can be processed by RF or QBE exactly as more classical visual feature vectors.

A simple solution for representing the set of keywords associated to the images as feature vectors consists in using one dimension for every keyword annotating an image. 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. Standard dimension reduction methods may provide more compact representations, but their quality is conditioned by the statistical representativity of the data. Also, the individual dimensions in these new representations would no longer be interpretable, so the individual feature vectors would not be comprehensible any more.

To obtain a scalable solution for representing sets of keywords as reliable and comprehensible feature vectors, we suggest to select a limited set of "core" concepts and to associate to every such concept a dimension in the feature vector. We rely on an external ontology, defining semantic relations between concepts, to find good candidates for the core concepts and to define the feature vectors for sets of keywords. WordNet is a well-known general purpose ontology that organizes nouns, verbs, adjectives and adverbs into synsets (set of words having similar meaning), each representing one underlying lexical concept. The concepts are linked by semantic relations of various types, such as synonymy, hypernymy, hyponymy, etc. Further details regarding WordNet can be found in [3], and [9].

The **core 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 in WordNet linking the concepts associated to all the keywords in the database to the most generic concepts. 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 (see Fig. 1 for an example subgraph). The paths obtained for all the keywords in the database define the hypernyms graph we will use later. A small set (compared to the number of different keywords) of core concepts is then manually selected;

good candidates are super-concepts of several $c(w)$ concepts that are relatively close to these; also, the core concepts must be balanced among all the branches containing $c(w)$ concepts. This issue is also under study in the text retrieval literature (e.g. [21]), but applied to data having different statistical characteristics. Then, we compute for every image a **conceptual feature vector** representing the "projection" of its keywords on the set of core concepts. We first study representations for a single keyword, then we turn to sets of keywords.

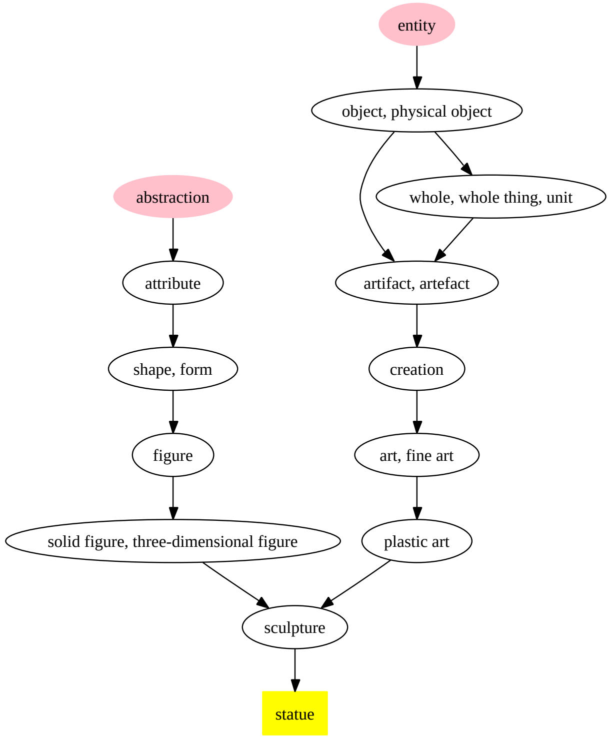

Figure 1: Hypernym subgraph generated by WordNet for the concept "statue".

In our feature vector, one dimension is dedicated to every core concept. Suppose that $\{C_i | 1 \leq i \leq n\}$ are the n core concepts selected. 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. One 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 keywords mapped to concepts that are different but have the same core super-concepts will have the same feature vectors. A refined solution should include in $\mathbf{v}(c(w))$ the *degrees of similarity* between $c(w)$ and its core super-concepts. Such a degree of similarity can be interpreted as the relevance of a core concept for describing an image annotated with the keyword. We thus have to evaluate the semantic similarity between concepts.

Apart from the binary measure presented above, there are several **measures of conceptual similarity**, relying on WordNet, that can be used for the definition of our keyword-based feature vector. The measures put forward in [14] and [25] rely on knowledge-rich sources (ontologies) alone, while those in [5], [16], [19] combine these sources with knowledge-poor sources (corpus statistics).

Leacock and Chodorow [14] rely on the length of the shortest path following IS-A relations, $\text{len}(c_1, c_2)$, 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: $\text{sim}_{LC}(c_1, c_2) = -\log(\text{len}(c_1, c_2)/2D)$. Wu and Palmer [25] evaluate the similarity according to how close the two concepts are in the concept hierarchy, $\text{sim}(c_1, c_2) = 2N_3/(N_1 + N_2 + 2N_3)$, where c_3 is 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 [19], the similarity between two concepts depends on the extent to which they share information. The similarity between two concepts is defined as the information content of their lowest super-ordinate $\text{lso}(c_1, c_2)$ according to $\text{sim}_R(c_1, c_2) = -\log p(\text{lso}(c_1, c_2))$, where $p(c)$ is the probability of encountering an instance of a concept c in some specific corpus. The proposal in Lin [16] is based on an information-theoretic similarity measure for arbitrary objects. With the notations above:

$$\text{sim}_L(c_1, c_2) = 2\log(p(\text{lso}(c_1, c_2)))/[\log(p(c_1)) + \log(p(c_2))]$$

In a comparative study, Budanitsky and Hirst [5] present the correlations between the human rating of similarity and several similarity measures. According to their results, among the measures described above, the Lin similarity measure is closest to the way human subjects interpret semantic similarity, which is why it is our main focus in the experimental evaluations.

Using any of these measures (indicated by the short names LCH, RES and respectively 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

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

In a second representation, we do not limit the evaluation of the similarity to the super-concepts of $c(w)$, so we set $v_i(c(w)) = \text{sim}(c(w), C_i)$; in the following, the use of this method will be indicated by the "-ALL" string 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 core concept only the keyword that is closest to this concept has an impact on the vector.

3. ACTIVE RELEVANCE FEEDBACK FRAMEWORK

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 [28]. Also, RF can be used for interactively defining image classes, so that annotations can be provided for an entire class at once (semi-automatic database annotation).

We use a SVM-based relevance feedback framework with two enhancements that can ameliorate its performances in an interactive image query scenario. To optimize the transfer of information between the user and the system, we employ a new active learning selection criterion that minimizes redundancy between the candidate images shown to the user. Also, we find that insensitivity to the spatial scale of the data is a desirable property for the SVM-machine employed as the learner in relevance feedback and we show how to obtain such insensitivity by the use of specific kernel functions. A detailed description of our RF scheme as well as extensive experimental evaluations, using several image databases, can be found in [10].

3.1 Reducing the redundancy

The concept of *semantic gap* is used in Content Based Image Retrieval (CBIR) community to qualify the difficulty associated with searching for semantic entities in image databases [12]. Relevance feedback is often used as a tool to approach the semantic gap by interactively asking the user to qualify the decisions made by the machine. 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.

Cox et al. [6] introduce some interesting ideas 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 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.

Tong et al. [24] present several selection criteria for SVM-learners applied to content-based image retrieval with relevance feedback. The simplest (and computationally cheapest) of these criteria consists in selecting the texts whose representations (in the feature space induced by the kernel) are closest to the hyperplane currently defined by the SVM. We call this simple criterion the selection of the "most ambiguous" (MA) candidate(s). This selection criterion is justified by the fact that knowledge of the label of such a candidate halves the version-space. While the MA criterion provides a computationally effective solution to the selection of the most ambiguous images (satisfying the first condition mentioned above), when used for the selection of more than one candidate image it does not remove the redundancies between the candidates.

In the early stages of the 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. For this situation, Mitra et al. [18] suggest a probabilistic framework where a sample is selected according to the likelihood that it belongs to the real set of support vectors. In the initial learning phase, when the actual set of support vectors is not close enough to the optimal set, their algorithm explores a higher number of interior points. 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.

For two candidates images, x_i and x_j, we require a low value for $K(x_i, x_j)$. If all the images of vectors in the input space have constant norm and if the kernel K is inducing a Hilbert structure on the feature space, then this condition corresponds to a requirement of quasi-orthogonality between the images in the feature space. We shall call this criterion the selection of the "most ambiguous and orthogonal" (MAO) candidates. 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 choose as a new example the vector $x_j \in S$ that minimizes the highest of the values taken by $K(x_i, x_j)$:

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

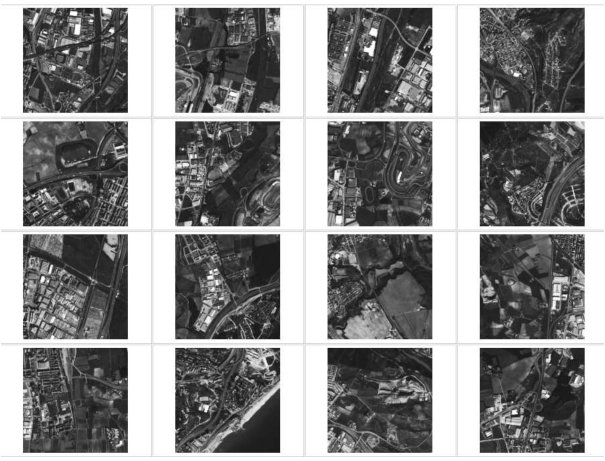

Figure 2: Example of RF-based retrieval in a database of 2800 satellite images. The user is looking for details of routes and highways, which are difficult to filter by a QBE approach. The images are returned by the system with 3 positive examples and 6 negative examples.

3.2 Invariance to scale

During the study of several groundtruth databases we found that the size of the various classes often covers an important 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 between parts of the frontier of some classes. A too strong sensitivity of the learner to the scale of the data could then strongly limit its applicability in an RF context. The kernels usually employed in SVM-based RF (e.g. the RBF kernel or the Laplace kernel) depend on a scale parameter that makes difficult to adapt to the scale of the data.

The triangular kernel, $K(x_i, x_j) = -\|x_i - x_j\|$, is a *conditionally* positive definite kernel, but the convergence of SVMs remains guaranteed with this type of kernel [20]. Fleuret and Sahbi [11] show that the triangular kernel have a very interesting property: it makes the frontier found by SVMs invariant to the scale of the data. In real applications, the scales of the user-defined classes cannot be known a priori and the scale parameter of a kernel cannot be adjusted online. The scale-invariance obtained by the use of the triangular kernel becomes then a highly desirable feature and experiments on several image databases prove this kernel to be a very good alternative.

4. EXPERIMENTAL EVALUATION

In this section we present experimental evaluation results of image retrieval using both the visual features we described and our new conceptual descriptor. We start by introducing the experimental setup and the performance measures we use, after which we present results both for QBE and relevance feedback.

4.1 Experimental setup

Ground truth database. We built our ground truth (GT) test database starting from an image database provided by Fratelli Alinari. This database has a heterogeneous content, featuring images illustrating many categories of human activity, e.g. art, archaeology, architecture, etc. There are 20000 images, 85601 annotations using 2059 keywords, many images being annotated by several keywords. We selected a test database having 3585 files for a total of 6664 annotations using 90 keywords. Keywords annotate between 26 and 274 images.

To have realistic classes, defined by both visual aspect and higher-level semantics, we built by hand a new ground truth, independent of the keyword annotations (no GT class is the union or intersection of sets of images annotated with the same keywords). We defined 20 classes in the GT, having between 15 and 174 images each. The number of images included in the groudtruth is 1073 and the degree of overlapping between classes is of about 10%. A certain degree of overlapping between GT 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 database, we perform all the evaluations on the entire database of 3585 images.

The conceptual feature vector. We built, as presented in Sec. 2.2, the hypernym graph associated with the whole test database and we choose 28 representative core concepts to be used for projecting the sets of keywords that annotate the images. Thus, the conceptual feature vector has 28 dimensions. No keyword was included as a core concept. To represent the visual content of the images, we use the visual feature vector presented in Sec. 2.1.

4.2 Evaluation of the relevance feedback mechanism

First, we evaluate the relevance feedback mechanism introduced in Sec. 3 on the test image database described above. 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 GT 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 GT class where the initial positive example belongs. When we use the MAO selection criterion, it is computed on a window of size $2 \times ws$.

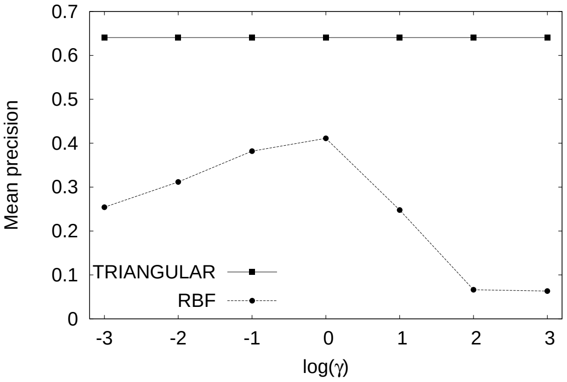

Figure 3: Mean precision vs. scale parameter for the RBF kernel and the triangular kernel.

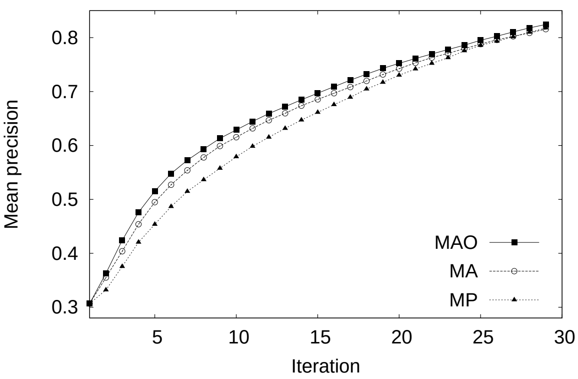

Figure 4: Comparison of several selection strategies for the triangular kernel using visual features and the WNS-LIN-ALL keyword-based signature.

We follow each relevance feedback session for 30 iterations (rounds)and we measure the precision within a window of size equal to the class size. 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 for each image in every class, at every iteration we compute the mean value of the precision measure over all feedback sessions. This provides a measure of how well performs relevance feedback, 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 introduced in Sec. 2.2.

First, we evaluate the sensitivity of the RBF kernel 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 iteration. This is a measure of how well performs relevance feedback with respect to the proposed GT for the given scale parameter. In Fig. 3 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 (that has no scale parameter). As we can see, the RBF kernel is very sensitive to the scale of the data. Moreover, no scale parameter value is really convenient for all classes in the ground truth, which explains why the performances of the RBF kernel are rather poor compared to those of the triangular kernel. The invariance to scale provided by the triangular kernel proves to be a very useful property for generalist databases, when the target class is complex and is best described in semantic terms.

Fig. 4 presents mean precision vs. iteration diagrams for several selection strategies: MAO, MA and MP (see Sec. 3). The MAO criterion provides better results than both MA and MP criteria. These new results both extend and confirm the evaluations presented in [10] for several GT databases and using visual descriptors alone.

4.3 Evaluation of the combined use of visual and conceptual descriptors

We present the evaluation of the combined use of visual features and the keyword-based WNS signatures both in a QBE context and with relevance feedback. For the **QBE** situation, we test several types of conceptual feature vectors (WNS signatures) presented in Sec. 2.2 and we build precision-recall diagrams using the ground truth described previously.

In Fig. 5 we present precision/recall diagrams for QBE using jointly the visual and WNS-LIN descriptors, the joint use of visual and WNS-LIN-ALL descriptors, and for the visual feature vector alone. The WNS-LIN-ALL signature performs clearly better than WNS-LIN when combined with the visual features, and much better than the visual feature vector alone. We obtained similar diagrams for the LCH and RES similarity measures. These findings were verified throughout the tests we performed in the QBE scenario: using both visual and conceptual feature vectors visibly improves the quality of the results compared to using visual features alone, and projecting the keywords on all the core concepts (WNS-LIN-ALL in the figure) gives better performance than projecting only on their core super-concepts. Projecting keywords on all the core concepts allows the use of semantic relations in WordNet other than hypernymy, through the similarity functions, which has a positive influence on the results returned by the system. However, we could not obtain experimental evidence to favor any of the similarity measures mentioned in Sec. 2.2.

We also tested the new WNS feature vectors using **relevance feedback** on our ground truth database.

In Fig. 6 we compare the WNS-BINARY signatures with WNS-LIN-ALL. We see that the LIN-ALL outperforms significantly the BINARY version, both when considered alone and when it is combined with the visual feature vector. Also, the joint use of visual and conceptual feature vectors considerably improves the results compared to the use of conceptual or visual features alone.

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 visual and conceptual feature vectors produces a dramatic improvement of the results, compared to the use of visual or conceptual descriptors alone.

The tests performed with relevance feedback strongly reinforce the conclusions of the QBVE evaluation: the conceptual descriptors relying on semantic similarity measures presented in Sec. 2.2 work better than the binary conceptual descriptor, and projecting the keywords on all the core concepts gives better results than using only the core super-concepts. Also, the joint use of both feature vectors performs much better than using the visual feature vector alone. Moreover, the improvements obtained by using the combined feature vector were much more visible with RF than in QBVE. 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 feature vectors, choosing during successive iterations only what is useful in the identification of the target.

As an illustration, in Figs. 8–10 we present three screens of results returned by our system in a QBE scenario, the query image being in the top-left corners of the screenshots.

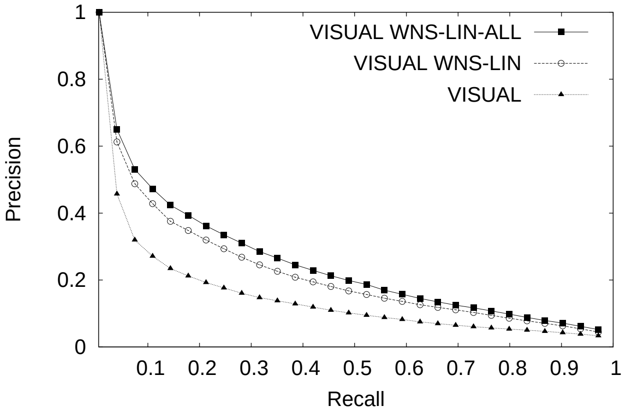

Figure 5: Precision/recall diagrams: the joint use of visual and conceptual feature performs significantly better compared to the use of visual features alone. Also, projecting the keywords on all the core concepts works better than projecting only on parent core concepts (WNS-LIN-ALL vs. WNS-LIN).

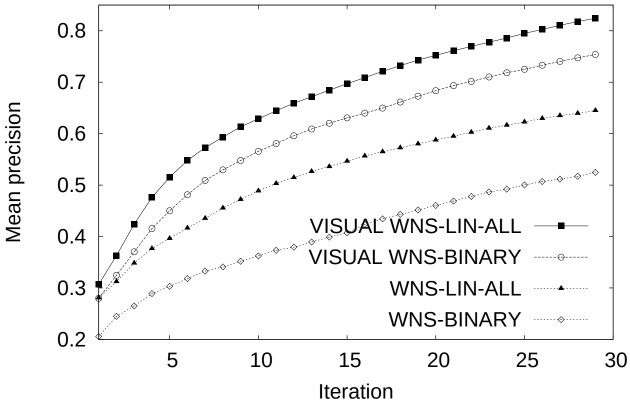

Figure 6: Relevance feedback: the combined use of visual and conceptual feature vectors give much better results compared to using the conceptual feature vector alone. Also, using a similarity measure (e.g. LIN) for the conceptual features provides better results than using the binary projection.

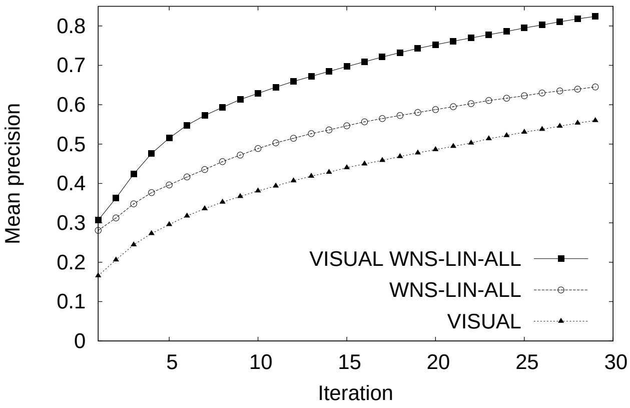

Figure 7: Relevance feedback: using the combined visual and conceptual descriptor dramatically improves the results compared to using the visual or conceptual descriptors alone.

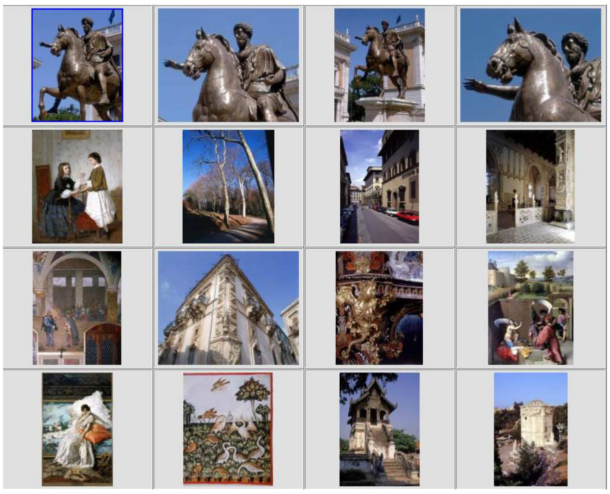

Figure 8: First page of QBE retrieval results with the visual descriptor.

In Fig. 8 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 the same visual appearance as the query image. The results in Fig. 9 correspond to the use of the WNS-LIN-ALL signature alone; while the returned images are conceptually related to the query image, their semantic content is too abstract and does not always represent well user's intent. Fig. 10 shows the results obtained when employing both visual and keyword-based descriptors. In this case, the returned images clearly correspond better to the intent of the user.

5. CONCLUSION

Although image retrieval using low-level visual features works well in many applications, in some situations the semantic gap limits its use with generic image databases. Alternatively, text annotations are more directly related to the high-level semantics of the images, but do not naturally reflect visual content. Keywords and visual features provide complementary information and using both of them is advantageous in many applications.

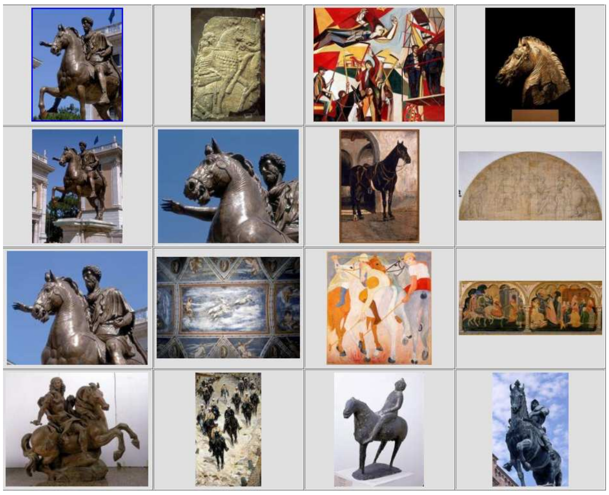

Figure 9: First page of QBE retrieval results with the WNS-LIN-ALL descriptor.

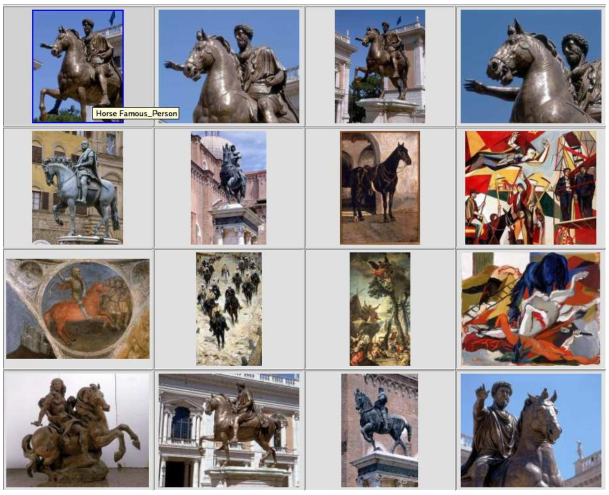

Figure 10: First page of QBE retrieval results with the combined visual and WNS-LIN-ALL descriptors.

In this paper, we introduced a new conceptual feature vector that makes use of an external ontology (WordNet) to induce a semantic generalization of the concepts corresponding to keywords. Evaluations performed on a ground truth build from a real world generalist database confirm that our new feature vector can improve dramatically the quality of the returned results, both with QBE and with relevance feedback. Moreover, because of its small memory use and computing time complexity, our new feature vector can be used with relevance feedback for large image databases.

6. ACKNOWLEDGEMENTS

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

7. 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] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. M. Blei, and M. I. Jordan. Matching words and pictures. *Journal of Machine Learning Research*, 3:1107–1135, March 2003.

[3] R. Beckwith, C. Fellbaum, D. Gross, and G. Miller. WordNet: A lexical database organized on psycholinguistic principles. In U. Zernik, editor, *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*, pages 211–232. Erlbaum, 1991.

[4] 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.

[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] 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.

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

[8] 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.

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

[10] 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.

[11] 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.

[12] T. Gevers and A. 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.

[13] 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.

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

[15] M. S. Lew. *Principles of Visual Information Retrieval.* Springer-Verlag, 2001.

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

[17] V. Mezaris, I. Kompatsiaris, and M. G. Strintzis. Region-based image retrieval using an object ontology and relevance feedback. *EURASIP Journal on Applied Signal Processing*, 2004(6):886–901, June 2004.

[18] P. Mitra and S. K. Pal. A probabilistic active support vector learning algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(3):413–418, 2004.

[19] 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.

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

[21] F. Seydoux and J.-C. Chappelier. Indexation sémantique au moyen de coupes de redondance minimale dans une ontologie. In *Proceedings of Traitement Automatique du Langage Naturel (TALN'05)*, pages 33–42, June 2005.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

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

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