Graph-Based Indexing and Querying on Image Corpora with Unified Visual Semantic and Relational Descriptions

Mohammed Belkhatir

School of Information Technology, Monash University
Belkhatir.mohammed@infotech.monash.edu

Abstract. We propose in this paper to integrate the semantic description of the image and the relational characterization of its components through an architecture which follows a sharp process for generating image index and query representations and computing their correspondence. This architecture relies on an expressive representation formalism handling high-level image descriptions and a conceptual query framework in an attempt to operate image indexing and retrieval operations beyond keyword-based and loosely-coupled state-of-the-art systems. At the experimental level, we evaluate its retrieval performance through recall and precision indicators on a test collection of 2500 color photographs.

1 Introduction and Related Work

State-of-the-art image indexing and retrieval systems are mainly based on characterizing the image content through an automatic process mapping low-level extracted features (such as color histograms or Gabor matrices for respectively color and texture extraction) to semantic-based keywords (among them [6,7,10]). The major disadvantage of this class of frameworks relies on the specification of restrained and fixed sets of semantic-based keywords which are moreover not sufficient to accurately represent non-textual documents, such as images. Regarding the fact that several artificial objects have high degrees of variability with respect to signal properties, an interesting solution is to extend the extracted visual semantics with signal characterizations in order to enrich the image indexing vocabulary and query language. Therefore, a new generation of systems integrating semantics and signal descriptions has emerged and the first solutions [8,13] are based on the association of textual annotations to characterize semantics with a relevance feedback (RF) scheme operating on low-level signal features. These approaches have three major drawbacks: first, they fail to exhibit a single framework unifying low-level data and semantics, which penalizes the performance of the system in terms of retrieval efficiency. Then, as far as the query process is concerned, the user is to query both textually in order to express high-level concepts and through several and time-consuming RF loops to complement his initial query. Therefore, this solution for integrating semantics and signal features, relying on a cumbersome query process, does not enforce facilitated and efficient user interaction. Finally, these systems do not take into account the relational spatial information between visual entities, which affects the quality of the retrieval results. Indeed, the need of an expressive index and query language for manipulating multimedia documents, and in particular one supporting relational characterization of the content, has been highlighted in [11].

D.H.-L. Goh et al. (Eds.): ICADL 2007, LNCS 4822, pp. 196–205, 2007.

We propose a unified multi-facetted framework unifying visual semantics and relational spatial characterization for automatic image retrieval that enforces expressivity through the use of symbolic descriptors to characterize the image content. After specifying a fully-automatic framework extracting the visual semantics, we enrich the description of images through the specification of processes establishing a correspondence between extracted low-level features and high-level spatial concepts. For example, with the semantic concepts "huts" and "grass" one might assign relations such as "above", "disconnected", "below" and "near" characterizing the fact that huts are above and disconnected from the grass, which is itself below and near huts. Therefore, not only do we characterize visual semantics, but also spatial relations linking them. For this, we consider an efficient operational model that allows relational indexing and is adaptable to symbolic image retrieval: **conceptual graphs** (CGs) [12]. However, contrarily to the EMIR^2 system [9] which was one of the early attempts at using CGs for image retrieval and limited its descriptive power to the basic semantics associated with these graphs (i.e. the conjunction of concepts and relations), we extend their operational semantics to handle a rich image query language consisting of the 3 major boolean operators (conjunction, disjunction and negation). Indeed, we are interested in dealing with non trivial queries involving the combination of visual semantics and spatial relations and the possibility to associate boolean operators to these queries. This would allow the user to retrieve images with "huts above **and** disconnected from the grass", "people at the left **or** at the right of buildings" or "houses **not** covered by vegetation"...

In the remainder, we first present the general organization of our image retrieval architecture. We deal in sections 3 and 4 with the visual semantics and spatial characterizations. Section 5 will specify the query framework. We finally present in section 6 the validation experiments conducted on a test collection of 2500 photographs.

2 An Architecture for Integrating Semantic and Spatial Descriptions

We propose an image retrieval architecture illustrated in fig. 1 which consists of five processing modules to integrate semantic and relational descriptions:

- The first provides the extraction of the image visual semantics through a statistical joint probability distribution tagging framework. Starting from a physical image (seen as a matrix of pixels), this framework allows to highlight the perceptually-meaningful visual entities with their associated semantic characterization in the form of a vector of semantic concepts with their recognition probabilities (further details are provided in section 3.1).
- The second module handles the image content representation and is based on a *multi-facetted* image model unifying visual semantics and spatial features. The **object facet** describes an image as a set of **image objects** (**IOs**) abstract structures representing visual entities within an image. The **visual semantics facet**, formally specified in section 3, describes the image semantic content and is based on labeling IOs with a semantic concept using the outcome of the semantic extraction module. E.g., in fig. 1, the first IO (Io1) is tagged by the semantic concept *People*. The **spatial facet**, detailed and formalized in section 4, describes the relational characterizations between pairs of IOs in terms of symbolic spatial relations. E.g. the first IO (Io1) is **inside** the second IO (Io2).

- The third module consists in applying the image model to specify an image index representation and therefore provides a representation of an image document in the corpus with respect to the multi-facetted image model. It is a structure called image index representation. In fig. 1, an image belonging to the corpus is characterized by a multi-facetted representation.
- At the other end of the architecture spectrum, the fourth module allows to translate a query with semantic and relational descriptions into a high-level structure with respect to the multi-facetted image model. It is a structure called image query representation. In fig. 1, a user full-text query "Images composed of people inside water" is translated into an image query representation closely following the multi-facetted image model.
- Finally, the fifth module deals with the correspondence process and the definition of a matching function between index and query representations. The image query representation is compared to all index representations of image documents in the corpus and a relevance value, estimating their degree of similarity is computed in order to rank all image documents relevant to a query. The search results are then displayed through the interface of the image retrieval system.

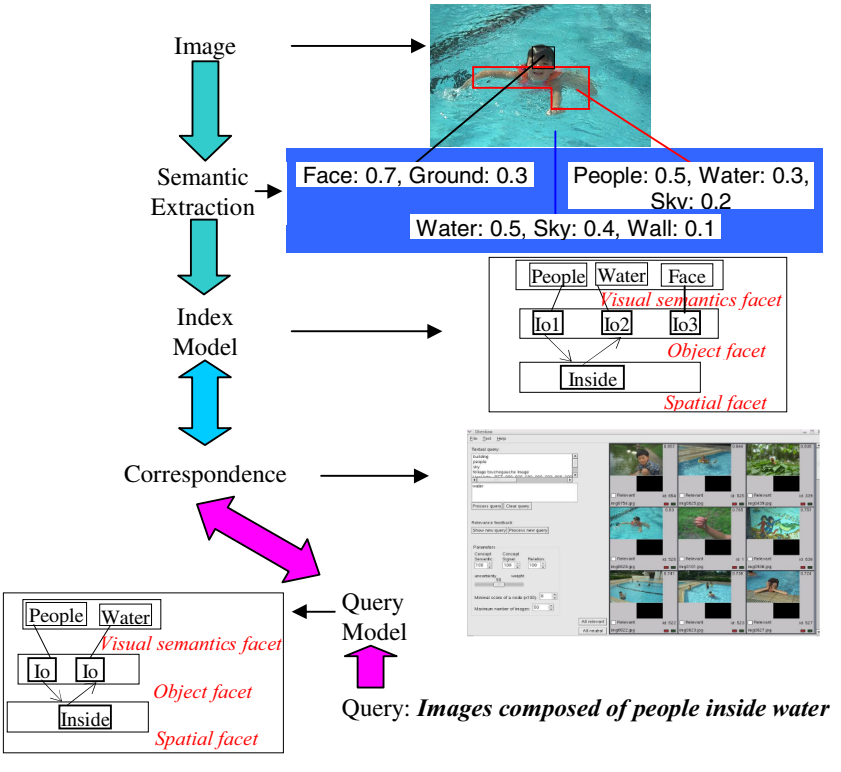

Fig. 1. Image retrieval architecture coupling semantic and relational descriptions

In order to instantiate the image model within an image retrieval framework, we choose a representation formalism capable to represent IOs, the visual semantics they convey and their relational characterizations: CGs. They have indeed proven to adapt to the symbolic approach of image retrieval [1,2,9,11] and allow to represent components of our image retrieval architecture and specify expressive index and query representations. Formally, a CG is a finite, bipartite, connex and oriented graph. It features two types of nodes: concept and relations. In the graph [2007]←(Year)←[ICADL]→(Location)→[Hanoi], concepts are between brackets and relations between parenthesis. This graph is semantically interpreted as: the ICADL conference of year 2007 is held in Hanoi. Concepts and relations are organized within lattice structures ordered by the IS-A relation.

3 The Visual Semantics Facet

3.1 Extracting the Semantics

Semantic concepts are learned and then automatically extracted given a visual ontology. Its specification is strongly constrained by the application domain [9]. Indeed dealing with corpus of medical images would entail the elaboration of a visual ontology that would be different from an ontology considering computer-generated images. In this paper, our experiments in section 6 are based on collections of general-purpose color photographs.

Several experimental studies presented in [10] have led to the specification of twenty categories or picture scenes describing the image content at a global level. Web-based image search engines (google, altavista) are queried by textual keywords corresponding to these picture scenes and 100 images are gathered for each query. These images are used to establish a list of semantic concepts characterizing objects that can be encountered in these scenes. A total of 72 semantic concepts to be learnt and automatically extracted are specified. Fig. 2 shows their typical appearance.

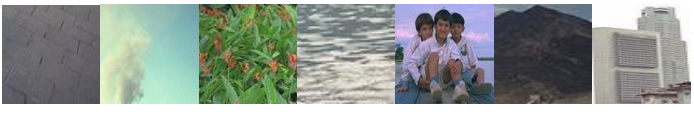

Fig. 2. Semantic concepts: ground, sky, vegetation, water, people, mountain, building

The indexing process is characterized by a statistical model which takes into account the joint distribution of semantic concepts on the one hand and symbolic signal features (color and texture) on the other hand. Starting from a learning set which includes IOs corresponding to visual entities, this model is instantiated by considering color and texture features of sets of connected rectangular regions used to generate the semantic concepts and their associated probabilities from this joint distribution. This process allows to highlight perceptually-meaningful visual entities with their associated semantic characterization in the form of a vector of semantic concepts with their recognition probabilities (further details can be found in [2]).

E.g., three visual entities linked to three IOs are highlighted from the example image in fig. 1. The first IO (*Io1*) is linked to a vector of semantic concepts with the

highest recognition probability corresponding to the concept *people*, the second IO (*Io2*) to a vector of semantic concepts with the highest recognition probability corresponding to the concept *water* and the third IO (*Io3*) to a vector of semantic concepts with the highest recognition probability corresponding to the concept *face.*

3.2 Model of the Visual Semantics Facet

IOs are represented by *Io* concepts and the semantic concepts are organized within a multi-layered lattice ordered by a specific/generic order. An instance of the visual semantics facet is represented by a set of CGs, each one containing an *Io* type linked through the conceptual relation *is_a* to a semantic concept. Let us note that only the semantic concept with the highest recognition probability is considered as far as the CG representation of the facet is concerned. The graph controlling the generation of all visual semantics facet graphs, called visual semantics graph, is: [Io]→(is_a)→[SC]. E.g., graphs [Io1]→(is_a)→[people], [Io2]→(is_a)→[water] and [Io3]→(is_a)→[face] are the representation of the visual semantics facet in fig. 1 and can be translated as: the first, second and third IOs are respectively associated to semantic concepts *people*, *water* and *face.*

4 The Spatial Facet: From Low-Level Spatial Features to High-Level Relational Description

Taking into account spatial relations between visual entities is crucial in the framework of an image retrieval system since it enriches the index structures and expands the query language. It is indeed shown in the study published in [5] that people frequently describe images by formulating spatial descriptions such as «…**at the left of**…» or «…**below**...». Also, dealing with relational information between image components allows to enhance the quality of the results of an information retrieval system [11]. We study in this part methods used to represent spatial data and deal with the automatic generation of high-level spatial relations following a first process of low-level extraction.

4.1 The Relation-Oriented Approach

In order to model the spatial data, we consider the «relation-oriented» approach which allows to explicitly represent the relevant spatial relations between IOs without taking into account their basic geometrical features. Our study features the four modeling and representation spaces:

- The Euclidean space gathers the coordinates of image pixels. Starting with this information, all knowledge related to the other representation spaces can be deduced.
- We consider in the topological space five relations inspired from [3] and justify this choice by the fact that they are exhaustive and relevant in the framework of an image retrieval system. Let io1 and io2 two IOs, these relations are (s_1=**P**,io1,io2) : 'io1 is a part of io2', (s_2=**T**,io1,io2) : 'io1 touches io2 (is externally connected)',

(s_3=**D**,io1,io2) : 'io1 is disconnected from io2', (s_4=**C**,io1,io2) : 'io1 partially covers (in front of) io2' and (s_5=**C_B**,io1,io2) : 'io1 is covered by (behind) io2'. Let us note that these relations are mutually exclusive and characterized by the the important property that each pair of IOs is linked by only one of these relations.

– The Vectorial space gathers the directional relations: Right (s_6=**R**), Left (s_7=**L**), Above (s_8=**A**) and Below (s_9=**B**). These relations are invariant to basic geometrical transformations such as translation and scaling.
– In the metric space, we consider the fuzzy relations Near (s_{10}=**N**) and Far (s_{11}=**F**).

4.2 Automatic Spatial Characterization

4.2.1 Topological Relations

In our spatial modeling, an IO *io* is characterized by its center of gravity io_c and by two pixel sets: its interior, noted io_i an dits border io_b. We define for an image an orthonormal axis with its origin being the image left superior border and the basic measure unity the pixel. All spatial characterizations of an object such as its border, interior and center of gravity are defined with respect to this axis.

In order to highlight topological relations between IOs, we consider the intersections of their interior and border pixel sets through a process adapted from [4]. Let io1 and io2 be two IOs, the four intersections are: io1_i ∩ io2_i, io1_i ∩ io2_b, io1_b ∩ io2_i and io1_b ∩ io2_b.

Each topological relation is linked to the results of these intersections as follows:

- (P, io1, io2) iff. io1_b∩io2_b = ∅, io1_i∩io2_b ≠ ∅, io1_b∩io2_i=∅ & io1_i∩io2_i ≠ ∅
- (T, io1, io2) iff. io1_b∩io2_b ≠ ∅, io1_i∩io2_b= ∅, io1_b∩io2_i=∅ & io1_i∩io2_i=∅
- (D, io1, io2) iff. io1_b∩io2_b=∅, io1_i∩io2_b=∅, io1_b∩io2_i=∅ & io1_i∩io2_i=∅
- (C,io1,io2) iff. io1_b∩io2_b=∅, io1_i∩io2_b=∅, io1_b∩io2_i ≠ ∅ & io1_i∩io2_i ≠ ∅
- (E_C, io1, io2) iff. io1_b∩io2_b=∅, io1_i∩io2_b ≠ ∅, io1_b∩io2_i=∅ & io1_i∩io2_i ≠ ∅

The strength of this computation method relies on associating topological relations to a range of necessary and sufficient conditions linked to spatial attributes of IOs (i.e. their interior and border pixel sets).

4.2.2 Directional Relations

The computation of directional relations between io1 and io2 is based on their centers of gravity io1_c($x1_c$, $y1_c$) and io2_c($x2_c$, $y2_c$), the minimal and maximal coordinates along x axis ($x1_{min}$, $x2_{min}$ et $x1_{max}$, $x2_{max}$) as well as the minimal and maximal coordinates along y axis ($y1_{min}$, $y2_{min}$ et $y1_{max}$, $y2_{max}$) of their four extremities.

We will say that io1 is at the left of io2, noted (L,io1,io2) iff. $x1_c < x2_c \wedge x1_{min} < x2_{min} \wedge x1_{max} < x2_{max}$. Also, io1 is at the right of io2, noted (R,io1,io2) iff. $x1_c > x2_c \wedge x1_{min} > x2_{min} \wedge x1_{max} > x2_{max}$.

We will say that io1 is above io2, noted (A,io1,io2) iff. $y1_c > y2_c \wedge y1_{min} > y2_{min} \wedge y1_{max} > y2_{max}$. Also, io1 is below io2, noted (B,io1,io2) iff. $y1_c < y2_c \wedge y1_{min} < y2_{min} \wedge y1_{max} < y2_{max}$.

We illustrate these definitions in fig. 3 where the IO corresponding to huts (io1) is above the IO corresponding to the grass (io2). It is however not at the left of the latter since $x1_c < x2_c$ but $x1_{min} > x2_{min}$.

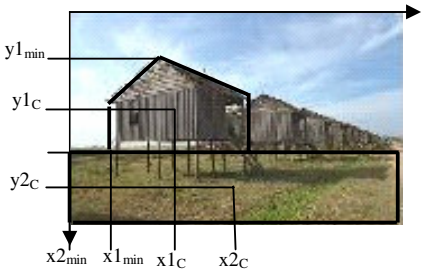

Fig. 3. Characterization of directional relations

4.2.3 Metric Relations

In order to distinguish between the Near and Far relations, we use the constant D_{sp}= d($\vec{0}$,0.5*$[\sigma_1,\sigma_2]^T$) where d is the Euclidean distance between the null vector $\vec{0}$ and $[\sigma_1,\sigma_2]^T$ is the vector of standard deviations of the localization of centers of gravity for each IO in each dimension from the overall spatial distribution of all IOs in the corpus. D_{sp} is therefore a measure of the spread of the distribution of centers of gravity of IOs. This distance agrees with results from psychophysics and can be interpreted as the bigger the spread, the larger the distances between centers of gravity are. Two IOs are **near** if the Euclidean distance between their centers of gravity is inferior to D_{sp}, **far** otherwise.

4.3 Conceptual Index and Query Structures for the Spatial Facet

4.3.1 Spatial Index Structures

IOs are related pairwise through an index spatial meta-relation (ISM), compact structure summarizing spatial relationships between these IOs. ISMs are supported by a vector structure **sp** with eleven elements corresponding to the previously explicited spatial relations. Values sp[i], i ∈ [1,11] are booleans stressing that the spatial relation s_i links the two considered IOs. E.g., Io1 is related to Io2 through the ISM <**P:1**, T:0, D:0, C:0, C_B:0, R:0, L:0, A:0, B:0, N:0, F:0>, translated as Io2 being part of (inside) Io3.

4.3.2 Spatial Query Structures

Our framework proposes an expressive query language which integrates visual semantics and symbolic spatial characterization through boolean operators. A user shall be able to link visual entities with a boolean conjunction of spatial relations such as in Q1: "huts **above** AND **disconnected** from grass", a boolean disjunction of spatial relations such as in Q2: "people **at the left** OR **at the right** of buildings" and a negation of spatial relations such as in Q3: "houses NOT **covered by** vegetation".

Three types of conceptual structures are specified to support the previously defined query types. *And* spatial meta-relations (ASMs) represent the signal distribution of an IO by a conjunction of spatial relations; *Or* spatial meta-relations (OSMs) by a disjunction of spatial relations and *Not* spatial meta-relations (NSMs) by a negation of spatial relations. The ASM <P:0, T:0, **D:1**, C:0, C_B:0, R:,0 L:0, **A:1**, B:0, N:0, F:0>$_{AND}$, the OSM <P:0, T:0, D:0, C:0, C_B:0, **R:1**, **L:1**, T:0, B:0, N:0, F:0>$_{OR}$ and the NSM <P:0,

T:0, D:0, C:0, **C_B:1**, R:0, L:0, T:0, B:0, N:0, F:0>$_{NOT}$ respectively correspond to the spatial characterizations featured in queries Q1, Q2 and Q3.

4.4 Graph Representation of the Spatial Facet

Spatial meta-relations are elements of partially-ordered lattices organized with respect to the type of the query processed (we will not detail this organization here). There are 2 types of basic graphs controlling the generation of all the spatial facet graphs. **Index spatial graphs** link two IOs through an ISM: **[Io1]→(ISM)→[Io2]**. **Query spatial graphs** link two IOs through *And*, *Or* or *Not* spatial meta-relations: **[Io1]→(ASM)→[Io2]**; **[Io1]→(OSM)→[Io2]** and **[Io1]→(NSM)→[Io2]**.

Eg, the index spatial graph [Io1]→[<**P:1**, T:0, D:0, C:0, C_B:0, R:0, L:0, A:0, B:0, N:0, F:0>]→[Io2] is a graph of the index representation of the spatial facet in figure 1 and is interpreted as: io1 is linked to io2 through the index spatial meta-relation <**P:1**, T:0, D:0, C:0, C_B:0, R:0, L:0, A:0, B:0, N:0, F:0> (i.e. io1 being part of io2).

5 The Query Module

Our conceptual architecture is based on a unified framework allowing a user to query with both semantic and relational descriptions. The representation of a query is, like image index representations, obtained through the combination (join operation [12]) of CGs over the visual semantics and spatial facets. E.g., the query Q1 is represented by the CG:

[Io1]→(sct)→[Huts] **[Io2]→(sct)→[Grass]**
[<P:0, T:0, **D:1**, C:0, C_B:0, R:,0 L:0, **A:1**, B:0, N:0, F:0>$_{AND}$**]**

Also, the query Q2 is represented by the CG:

[Io1]→(sct)→[People] **[Io2]→(sct)→[Buildings]**
[<P:0, T:0, D:0, C:0, C_B:0, **R:1**, **L:1**, T:0, B:0, N:0, F:0>$_{OR}$**]**

Finally, the query Q3: "houses not covered by vegetation" is represented by the CG:

[Io1]→(sct)→[Houses] **[Io2]→(sct)→[Vegetation]**
[<P:0, T:0, D:0, C:0, **C_B:1**, R:0, L:0, T:0, B:0, N:0, F:0>$_{NOT}$**]**

The evaluation of similarity between index and query representations is achieved through a correspondence function: the CG projection operator. This operator allows to identify within the index CG sub-graphs with the same structure as the query CG, with nodes being possibly restricted (i.e. they are specializations of the query CG nodes).

6 Validation Experiments: An Application to Home Photographs

We implement the theoretical framework presented in this paper and validation experiments are carried out on a corpus of 2500 color photographs used as a validation corpus in [1,2,7]. IOs within the 2500 photographs are automatically assigned a vector of semantic concepts with their corresponding recognition probabilities as shown in section 3.1 and characterized with a visual semantics facet CG as shown in section 3.2. Also, pairs of IO are characterized with spatial index structures (section 4.3.1) and linked through an index spatial graph as shown in section 4.4.

As opposed to state-of-the-art keyword-based frameworks [6,7,10], we wish to retrieve photographs that represent elaborate image scenes and propose 12 queries

characterizing relative location of visual entities such as *people near vegetation* along with their ground truths among the 2500 photographs. The evaluation of our formalism is based on the notion of *image relevance* which consists in quantifying the correspondence between index and query images. We compare our system with a system based on a semantic keyword-based approach: the *Visual Keyword* system S_1 [7] and a state-of-the-art loosely-coupled system S_2 combining a textual framework for querying on semantics and a RF process operating on low-level signal features.

For each proposed query in table 1, we construct relevant textual query terms using corresponding visual semantics and spatial characterizations as input to our system (e.g. *people near vegetation*). The retrieval results for this query are given in fig. 4. S_1 processes three series of three random relevant photographs for each query (they correspond to people near vegetation as far as our example query is concerned). Also each query in table 1 is translated in relevant textual data to be processed by the semantic framework of S_2 ('people, vegetation' for *people near foliage*). Then to refine the results, three random relevant photographs are selected as input to the RF framework.

Fig. 4. Results for "People near vegetation"

Table 1. Queries

People touch pool
Buildings left and right of people
People in (part of) water
Foliage left and right of people
People near buildings
People in front of mountains
Close-up of people (people not related to any IO)
Close-up of buildings (buildings not related to any IO)
People in front of buildings
People near foliage
Cityscape (view from far)
Mountain (view from far)

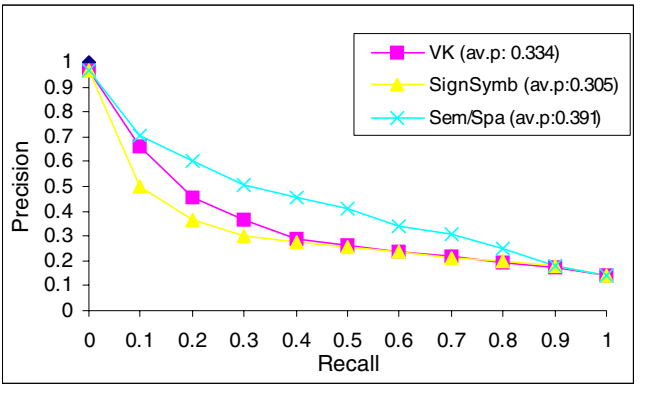

Fig. 5. Recall/Precision curves

Recall/precision curves of fig. 5 illustrate the average results obtained for all queries considering the corpus of 2500 images: the curve associated with the *Sem/Spa* legend illustrates the results in recall and precision obtained by our system, the curve associated with the *VK* legend by S_1 and the curve associated with the *SignSymb* legend by S_2. The average precision of our system (0.391) is approximately 17,1% higher over the average precision of the VK system (0.334) and approximately 28,2% higher over the average precision of the loosely-coupled state-of-the-art system (0.305). We notice that improvements of the precision values are significant at all recall values. This shows that when dealing with elaborate queries which combine multiple sources of information (here visual semantics and spatial characterizations) and thus require a higher level of abstraction, the use of an "intelligent" and expressive representation formalism (here the CG formalism within our framework) is crucial. As a matter of fact, our system complements automatic keyword-based approaches (in this case the VKs) through the enrichment of their query frameworks with spatial characterization. Moreover, it outperforms state-of-the-art loosely-coupled solutions by proposing a unified high-level and expressive framework optimizing user interaction and allowing to query with precision over visual semantics and symbolic spatial relations.

References

1. Belkhatir, M.: A Full-Text Framework for the Image Retrieval Signal/Semantic Integration. In: Andersen, K.V., Debenham, J., Wagner, R. (eds.) DEXA 2005. LNCS, vol. 3588, pp. 113–123. Springer, Heidelberg (2005)
2. Belkhatir, M.: On the Signal/Semantic Integration for Symbolic Image Indexing and Retrieval. PhD Thesis of the Joseph Fourier University (2005)
3. Cohn, A., et al.: Qualitative Spatial Representation and Reasoning with the Region Connection Calculus. Geoinformatica 1, 1–44 (1997)
4. Egenhofer, M.: Reasoning about binary topological relations. In: Proceedings of SSD, pp. 143–160 (1991)
5. Hollink, L., et al.: Classification of user image descriptions. Int. Journal of Human Computer Studies 61(5), 601–626 (2004)
6. Jeon, J., et al.: Automatic image annotation and retrieval using cross-media relevance models. In: Proceedings of ACM SIGIR, pp. 119–126 (2003)
7. Lim, J.H., Jin, J.S.: A structured learning framework for content-based image indexing and visual query. Multimedia Systems 10(4), 317–331 (2005)
8. Lu, Y., et al.: A unified framework for semantics and feature based relevance feedback in image retrieval systems. In: Proceedings of ACM Multimedia, pp. 31–37 (2000)
9. Mechkour, M.: EMIR2: An Extended Model for Image Representation and Retrieval. In: Revell, N., Tjoa, A.M. (eds.) DEXA 1995. LNCS, vol. 978, pp. 395–404. Springer, Heidelberg (1995)
10. Mojsilovic, A., Rogowitz, B.: Capturing image semantics with low-level descriptors. In: Proceedings of IEEE ICIP, pp. 18–21 (2001)
11. Ounis, I., Pasca, M.: RELIEF: Combining expressiveness and rapidity into a single system. In: Proceedings of ACM SIGIR, pp. 266–274 (1998)
12. Sowa, J.F.: Conceptual structures: information processing in mind and machine. Addison-Wesley publishing company, Reading (1984)
13. Zhou, X.S., Huang, T.S.: Unifying Keywords and Visual Contents in Image Retrieval. IEEE Multimedia 9(2), 23–33 (2002)