Multimedia Systems (2011) 17:135–148
DOI 10.1007/s00530-010-0207-8

ORIGINAL RESEARCH

A three-level architecture for bridging the image semantic gap

Mohammed Belkhatir

Received: 5 December 2008 / Accepted: 10 October 2010 / Published online: 16 November 2010

Abstract Image retrieval systems face the problem of dealing with the different ways to apprehend the content of images and in particular the difficulty to characterize the visual semantics. To address this issue, we examine the use of three abstract levels of representation, namely Signal, Object and Semantic. At the Signal Level, we propose a framework mapping the extracted low-level features to symbolic signal descriptors. The Object Level features a statistical model considering the joint distribution of object concepts (such as *mountains*, *sky*…) and the symbolic signal descriptors. At the Semantic Level, signal and object characterizations are coupled within a logic-based framework. The latter is instantiated by a knowledge representation formalism allowing to define an expressive query language consisting of several boolean and quantification operators. Our architecture therefore makes it possible to process topic-based queries. Experimentally, we evaluate our theoretical proposition on a corpus of real-world photographs and the TRECVid corpus.

Keywords Multimedia processing · Semantic gap · Image indexing and retrieval · Experimental evaluation

1 Introduction

Image indexing and retrieval systems, which have been the subject of extensive research works since the 1990s, can be categorized with respect to their index and query abstraction level. We mainly identify three levels:

The first level, namely *Signal Level*, represents numerical abstractions of image regions. Such abstractions characterize the colors, textures… of visible elements in images. The general approach consists in computing structures representing the image distribution such as color histograms, texture features and using this data to partition the image; thus reducing the search space during the image retrieval operation. These methods hold the advantage of being fully automatic, thus are able to quickly process queries. Aspects related to human perception, which are of prime importance in image retrieval, are however not taken into account. In the remainder of the paper, this level is considered only as far as the automatic extraction of low-level signal features is concerned.

In order to address the impossibility of the signal-based systems to characterize the image semantics (also called semantic gap [1]), the second level (namely *Image Object Level*) of representation supports the notion of labeling the image visual entities. This level intends to bridge a gap between the signal aspects (first level) and the symbols representing the content of images. For this, two classes of automatic semantic extraction architectures have been proposed in the literature. The first, which aims at categorizing images in broad semantic classes, operates at the global image level. In [2], several experimental studies lead to the specification of 20 semantic categories or image scenes describing the image content at a global level (such as *group of people*, *cityscapes*, *landscapes*…). Each of these categories is then linked to several low-level features gathered within the complete feature set. The most recent automatic annotation models linking annotation words to visual features are based on statistical models [3–8]. Blei and Jordan [3] extend Dirichlet's latent allocation model

Communicated by Wei-Ying Ma.

M. Belkhatir (✉)
CNRS, University of Lyon, Lyon, France
e-mail: mohammed.belkhatir@iut.univ-lyon1.fr

and propose a correlation model linking words and images. The latter is based on the hypothesis that a combination of latent factors can be sampled from a Dirichlet distribution and used to generate words and image regions. This parametric model is based on the Expectation–Maximization algorithm to estimate these factors. Frameworks which have shown interesting performance improvements [5, 6, 8] are based on a doubly non-parametric approach for which the probabilities of associating words to image features are learnt from each image of a training set. They are then used to generate the probability of linking a word to a given query image. In [4], the image annotation task is modeled by a supervised multi-class labeling architecture. In [9, 10], the process of image indexing is enriched through the use of external linguistic knowledge bases (e.g., WordNet). The approach is guided by the dependencies between annotating words represented by a hierarchy derived from a textual ontology. While the above models predict the probability of an annotation word given an image, one is interested in generating the set of all index words representing all the image visual entities.

A second class of architectures, operating at the visual entity level has been proposed in [11–13]. One of the early solutions presented a probabilistic framework based on estimating class likelihoods of local areas, labeled as either man-made versus natural or inside versus outside objects [11]. In [13], training sample of the image regions is categorized into 11 clusters through a neural network mapping (e.g., *tree*, *fur*, *sand*…). To alleviate the restrained cardinality of the proposed previous sets of visual clusters, a richer index vocabulary consisting of 26 image labels called Visual Keywords (such as *sky*, *people*, *water*…) is specified in [12, 14]. However, this solution relies on a query-by-example solution for querying and no language being able to manipulate the extracted semantics has been proposed. The main disadvantage of this second class of frameworks relies on the specification of restrained and fixed sets of semantic classes.

The third level (namely "*Semantic Level*") is dedicated to represent the explicit characteristics of the image objects of the second level. Regarding the fact that several artificial objects have high degrees of variability with respect to signal properties such as color and texture variations, an interesting solution is to extend the extracted object-based descriptions with signal characterizations in order to enrich the image indexing vocabulary and query language, e.g., with the object concept "sky" one might assign additional concepts such as "cyan", "grey" characterizing its color and "covered", "smooth" which feature its texture. The third level representation is based on expressive representation formalisms able to support logical inference, for instance specialization/generalization (Is_A) hierarchies of concepts in a way to extend the retrieval capabilities. A class of frameworks within the European Fermi project implement this third level by proposing to model the image semantic and signal contents following a sharp process of human-assisted indexing [15, 16]. These approaches, based on elaborate knowledge-based representation models, provide satisfactory results in terms of retrieval quality but are not easily usable on large collections of images because of the necessary human intervention required for indexing.

If we examine a notion of abstraction targeted at each level (left part of Fig. 1), the Signal Level corresponds to low abstraction (because it considers low-level features), the Image Object Level is a medium abstraction representation (since it abstracts the signal to object-based characterization but does not go further), the Semantic Level is a high abstraction level because it features further characterization of the nature of objects and allows interpretation of scenes. The structure of Fig. 1 shows that the three levels of indexes are obviously not independent from each other. However, the grey parts that represent the transition between the levels are far from being easy tasks: going from signal features to visual objects usually relies on some learning process, whereas going from Object to Semantic Levels needs human input in state-of-the-art systems.

The right part of Fig. 1 highlights the querying expressiveness. It varies from low expressiveness at the Signal Level where the user inputs one or more query images through query-by-example (QBE) or relevance feedback (RF) to high expressiveness at the Semantic Level where the user is able to formulate a query through natural language interaction.

We contribute in this paper to the enrichment of the three levels of representation and highlight steps to bridge the gaps between them.

We enhance the Signal Level through the specification of processes establishing a correspondence between extracted low-level features and high-level visual information. For this, we specify a learning agent-based framework categorizing signal color and texture low-level features into symbolic categories.

We contribute to the Image Object Level through an automatic object-based indexing framework, operating at the visual entity level and characterized by a statistical

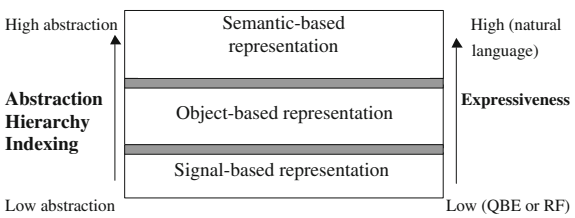

Fig. 1 Index and query representation levels

model which takes into account the joint distribution of object-based concepts on the one hand and color and texture categories on the other hand.

At the Semantic Level, the symbolic signal information and object characterizations are coupled through a unified model to enrich the image description. This model is instantiated by a logic-based knowledge representation formalism enforcing expressiveness in order to expand the index and query languages. As far as the interaction is concerned, since users are more skilled in defining their information needs using language-based descriptors [14], query formulation relies on natural language interaction with the possibility to manipulate several boolean and quantification operators. We are therefore able to process not only single-concept queries such as the automatic semantic extraction architectures, but also non-trivial queries involving multiple characterizations such as proposed in the framework of the TRECVid topic search track: visual semantics and high-level color concepts ("a *grey* sky"), visual semantics and high-level texture concepts ("fields of *lined* flowers").

In the remainder, we propose in Sect. 2 the characterization of the Signal Level. Section 3 details the object-based indexing framework (level 2). Section 4 deals with the Semantic Level and the coupling of the signal and object-based characterizations. Section 5 details the query module. Section 6 presents our experimental evaluation on both a corpus of 2,500 photographs and the TRECVid corpus.

2 Enhancing the Signal Level through mapping low-level features to signal categories

Our framework uses learning agents which have the ability to perceive color or texture signals and categorize their perception as symbolic color or texture categories.

2.1 Color perception and representation

When perceiving the physical world, a mapping is made from the physical space to a representation in the psychophysical space. Upon this representation, further cognitive actions such as categorization or recognition are taken. The representation should fulfill three requirements. First, it should be a good model for how humans perceive color. Second, it should make discrimination possible: two stimuli are discriminable if and only if they map onto different points in the representation space. Third, one should be able to define a similarity measure over the representation space. The HVC perceptive color space satisfies these requirements and has proven its merit at categorizing symbolic colors [22]. It belongs to the category of user-oriented color spaces (as opposed to material-oriented spaces such as RGB), i.e., spaces which define color as being perceived by a human through tonality (describing the color wavelength), saturation (characterizing the quantity of white light in the color spectral composition) and brightness (related to color intensity).

Our symbolic representation of color information is guided by the research carried out in color naming and categorization. Under the impulsion of Berlin and Kay [17], works have revolved around stressing a step of correspondence between color stimuli and 'basic color terms' which they characterize by the following properties: their application is not restricted to a given object class, i.e., the color characterized by the term "olive color" is not valid; they cannot be interpreted conjointly with object parts, i.e. "the maple leaf color" is not a valid color; their interpretation does not overlap with the interpretation of other color terms and finally they are psychologically meaningful. Given a series of perceptive evaluations and observations, 11 color categories ($c_i \in C_{\text{cat}}$) are highlighted: c_1 = cyan (C), c_2 = white (W), c_3 = green (Gn), c_4 = grey (G), c_5 = red (R), c_6 = yellow (Y), c_7 = black (B), c_8 = blue (Bl), c_9 = orange (O), c_{10} = purple (P), c_{11} = skin (S).

2.2 Texture perception and representation

The study of texture in computer vision has lead to the development of several computational models for texture analysis used in several CBIR architectures [1]. However, these texture extraction frameworks mostly fail to capture aspects related to human perception. Therefore, we propose a solution specifying a computational framework for texture extraction which is the closest approximation of the human visual system. The action of the visual cortex, where an object is decomposed into several primitives by the filtering of cortical neurons sensitive to several frequencies and orientations of the stimuli, is simulated by a bank of Gabor filters. An object is characterized by its Gabor energy distribution within seven spatial frequencies covering the whole spectral domain and seven angular orientations. This constitutes a texture space consisting of 49-dimension vectors, each dimension corresponding to a Gabor energy.

Our symbolic representation of texture information is guided by the texture lexicon proposed in [18] consisting of 11 high-level texture categories as a basis for symbolic texture classification. In each of these categories, several texture words which best describe the nature of the characterized texture are proposed. We consider the following texture categories ($t_i \in T_{\text{cat}}$): t_1 = bumpy (B), t_2 = cracked (C), t_3 = disordered (D), t_4 = interlaced (I), t_5 = lined

(L), t_6 = marbled (M), t_7 = netlike (N), t_8 = smeared (S), t_9 = spotted (Sp), t_{10} = uniform (U) and t_{11} = whirly (W).

2.3 Color and texture categorization

When an agent is to communicate about the world, a symbolic representation of the perception is needed. This symbolic representation arises by cutting up and structuring the representation space.

The color and texture spaces are used to define categories. A category has a number of features (the tonality, saturation and brightness values for colors and the Gabor energy dimensions for textures) and for each feature a fuzzy membership function is defined. If an unknown stimulus is perceived, a measure is needed of how well a category matches the unknown representation. To represent a category, a radial basis function with one output unit, divided by the number of hidden units, is chosen. It is our preferred choice for representing categories since it can divide the input space into regions whose configuration is not restricted in any way. A second advantage is that it is easily analyzed, which is valuable for monitoring the performance of the categorization.

Figure 2 shows the radial basis function. It consists of a layer of an unspecified number of hidden units acting as tuned receptors and one output unit. The input x is a perceptive color representation, i.e. a three-dimensional vector containing the tonality, saturation and brightness values, or a perceptive texture representation consisting of the Gabor energy dimensions. The hidden units are Gaussian functions $z_j(x)$. The output of the network $y(x)$ is the weighted sum of the Gaussians, weighted by the number of hidden units.

The goal of our framework is to successfully distinguish color or texture stimuli as being related to any color or texture categories. It follows a simple algorithm, and is completed by a learning agent with a set of color or texture categories. A random context $S = \{s_1,\ldots, s_n\}$ is created and presented to the agent. It contains n color or texture stimuli s_i of which one stimulus is the topic. The topic has to be discriminated from the rest of the context. The process is as follows:

1. Context $S = \{s_1,\ldots, s_n\}$ and the stimulus s_t are presented to the learning agent.
2. The learning agent perceives each stimulus s_i and returns a perceptive representation for each stimulus: $P(s_i) = \{s_1(o_1),\ldots, s_m(o_i)\}$.
3. For all n perceptive representations, the closest matching category c_s is found. $\forall c \in C_{cat}$ or T_{cat}: $y_c(P) \leq y_{c\_s}(P)$ where y_c is the output of the adaptive network belonging to category c, and y_{c_p} is the output of the adaptive network reacting best to P.
4. The stimulus s_t can be discriminated from the context when there exists a category matching the topic but not matching any other stimuli in the context.
5. We associate a confidence value for the categorization. For this, we use the distance to the decision boundary $f_t(c)$ (where f_t is the trained adaptive network related to category c) and map it on *posterior recognition probabilities*. In order to achieve this mapping, a logistic classifier maximizing the likelihood of the categorization is used.

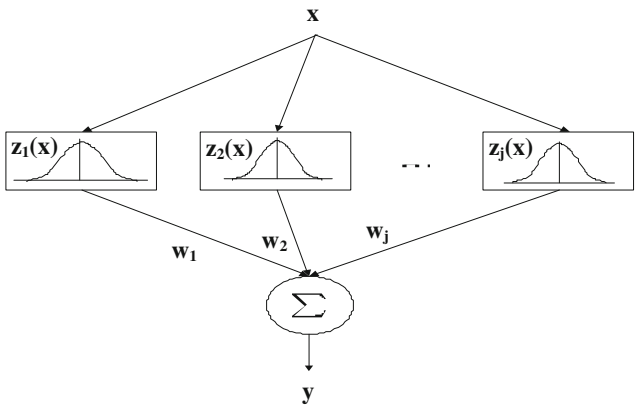

Fig. 2 Radial basis function for representing color or texture categories

3 A framework for highlighting image objects mapping signal categories to object concepts

At the core of our Object Level architecture is the notion of image objects (IOs), abstract structures representing visual entities within an image. Their specification is an attempt to operate image indexing and retrieval operations beyond simple low-level processes [1]. We consider a training set T comprised of annotated IOs. Let an IO io_t within this set, it is represented by a set of rectangular regions $r_{io\_t} = \{r_{1\_t},\ldots, r_{n\_t}\}$ and is indexed by an object concept c_{obj}, sets of color categories $\{c_{col\_1},\ldots, c_{col\_n}\}$ and texture categories $\{c_{tex\_1},\ldots, c_{tex\_m}\}$ (where $m, n \leq 11$)

3.1 Formal model

In the framework of the object-based indexing, we consider applying on a new image (i.e., not indexed) a rectangular grid defining **nb_grid** rectangular regions. Let us consider a set **rc** comprised of one to nb_grid connected rectangular regions. The reader shall note that this set of connected rectangular regions does not forcibly define an IO (we explicit in Sect. 3.2 the conditions for defining an IO).

We consider the existence of a probability distribution $P(.|rc)$ which can be seem as a finite set containing all object concepts which possibly index a set of rectangular regions as well as the color and texture categories characterizing it with their associated probabilities.

The set rc is defined in terms of object concepts $c_{obj}[1]\ldots c_{obj}[i]$ and color and texture categories $\{c_{col\_rc}[1],\ldots, c_{col\_rc}[j], c_{tex\_rc}[1],\ldots, c_{tex\_rc}[k]\}$, $i \leq \text{Card}_{obj}$, $j \leq 11$ and $k \leq 11$ (Card_{obj} is the maximum cardinality of the set of object concepts). We would like to index rc with an object concept. For this, we extract a concept c_{obj_rc} from its probability distribution $P(.|rc)$ such that $P(c_{obj\_rc}|rc) = \max[P(c_{obj}[i]|rc)]$, $i \leq \text{Card}_{obj}$. We therefore need to estimate the probability $P(c_{obj}[i]|rc)$ for each specified object concept. Given that $P(.|rc)$ is unknown, the probability of extracting the object concept $c_{obj}[i]$ is approximated by the conditional probability $P(c_{obj}[i]|c_{col\_rc}[1],\ldots, c_{col\_rc}[j], c_{tex\_rc}[1],\ldots, c_{tex\_rc}[k])$. We assume that the set of connected rectangular regions rc (possibly unique) is generated with respect to an undetermined distribution conditioned by the set $\{c_{col\_rc}[1],\ldots, c_{col\_rc}[j], c_{tex\_rc}[1],\ldots, c_{tex\_rc}[k]\}$. We will make no attempt to model the generation process of rc from it. The connected rectangular regions are gathered to constitute rc.

Considering the training set T of annotated IOs, we estimate the joint probability of observing the object concept $c_{obj}[i]$ with color categories $c_{col\_rc}[1],\ldots, c_{col\_rc}[j]$ and texture categories $c_{tex\_rc}[1],\ldots, c_{tex\_rc}[k]$. We take as a hypothesis that the observation of $c_{obj}[i]$ and $c_{col\_rc}[1],\ldots, c_{col\_rc}[j], c_{tex\_rc}[1],\ldots, c_{tex\_rc}[k]$ are independent events, a direct consequence of modeling connected regions by a finite set containing all object concepts which can possibly index these connected regions and all color and texture categories. We operate the marginalization of the distribution with respect to $c_{obj}[i]$, which is then determined as an expectation over all IOs in T:

$$P\left(c_{obj}[i]|c_{col\_rc}[1],\ldots, c_{col\_rc}[j], c_{tex\_rc}[1],\ldots, c_{tex\_rc}[k]\right) = \frac{\sum_{io\_t\in T}[P(\text{io\_t})P(c_{sem}[i]|\text{io\_t})(P(c_{col\_rc}[1]|\text{io\_t})\ldots P(c_{col\_rc}[j]|\text{io\_t}))(P(c_{tex\_rc}[1]|io\_t)\ldots P(c_{tex\_rc}[k]|\text{io\_t}))]}{\sum_{c_{sem}[l],l\in[1,\text{Card}_{obj}]} P(c_{sem}[l], c_{col\_rc}[1],\ldots, c_{col\_rc}[j], c_{tex\_rc}[1],\ldots, c_{tex\_rc}[k])} \quad (1)$$

Probabilities $P(\text{io\_t})$ are uniform considering all IOs of the set T. To evaluate the probability $P(c_{obj}[i]|\text{io\_t})$, we use maximum probability estimates.

Probabilities $P(c_{col\_rc}[1] \mid \text{io\_t})\ldots P(c_{col\_rc}[j] \mid \text{io\_t})$ and $P(c_{tex\_rc}[1] \mid \text{io\_t})\ldots P(c_{tex\_rc}[k] \mid \text{io\_t})$ are given by the processes, respectively, linking each color category $c_{col\_rc}[1],\ldots, c_{col\_rc}[j]$ to its probability, i.e. the percentage of pixels belonging to the corresponding set of connected rectangular regions (cf. Sect. 2.1) and each texture category $c_{tex\_rc}[1],\ldots, c_{tex\_rc}[k]$ to its posterior recognition probability through the learning framework applied to the considered set of connected rectangular regions (cf. Sect. 2.3).

3.2 Application

Starting from a physical image (i.e., non-indexed), we apply a fixed-size rectangular grid to subdivide it in rectangular regions $r_1,\ldots, r_{nb\_grid}$. Then we determine the color and texture categories linked to the sets of connected regions and their correlated probabilities (cf. Sects. 2.1, 2.2). We then use Eq. (1) to determine the object concept with the highest probability to co-occur with the color and texture categories within the set of connected regions.

When the object concepts with maximum probability linked to all sets comprised of 1 to nb_{reg_grid} connected regions are determined, we define image objects as the sets of connected regions with highest cardinality in which the object concept with highest probability is the same as those of their subsets of connected regions. Results are then reconciled across sets of adjacent rectangular regions to highlight image objects and their associated most probable object concept.

Regarding the sets of connected regions with a zero probability for all object concepts, we consider as an image object the set of maximum cardinality for which all subsets are composed of object concepts with zero probabilities. This image object is indexed by the object concept *unknown*.

Let us note that blocks which do not correspond to any object concept (i.e., the recognition probability for all object concepts is null) are still taken into account and aggregated. Although they do not convey any visual object-based information, they will be characterized by symbolic color, texture and relational information directly exploitable for query composition.

4 Coupling signal and object concepts within the Semantic Level

The integration of signal and object-based information within the retrieval framework is crucial since it expands the query language with the possibility to query over both object-based and visual information. At the Semantic Level, we propose an image model considering an image as

a multi-facetted object with the two principal facets being the physical (considering an image as a matrix of pixels) and the *logical* facets. The logical facet, grouping all aspects of the image content and its general context, is itself an aggregation of two basic facets: the object and signal facets. It characterizes the index and query image contents through a first-order logic formula.

- The *object facet* describes the image semantic content and is based on labeling image objects with an object concept. This facet gathers the sets C_{obj} of object concepts.
- The *signal facet* describes the image signal content in terms of symbolic perceptive features and consists in characterizing image objects with signal concepts. It itself consists of two subfacets.

The *color subfacet* features the image signal content in terms of color categories. This facet gathers the sets C_{col} of color concepts. The *texture subfacet* describes the signal content in terms of symbolic texture features. This facet gathers the sets C_{tx} of texture concepts.

In order to instantiate this model as an image retrieval framework, we shall consider a representation formalism well-suited for our logical formulation and capable of representing image objects as well as the visual semantics and signal information they convey. Moreover, this representation formalism should provide an intelligible representation of the information related to an image. It should therefore combine expressiveness and a user-friendly representation. Meta-concepts are an efficient solution to describe an image and characterize its components. The asset of this knowledge representation formalism is its flexible adaptation to the symbolic approach of image retrieval and allows uniformly representing components of our architecture and developing expressive and efficient index and query frameworks. Formally a meta-concept is represented by a vector-like syntactical structure with each dimension corresponding to a value (boolean or quantified) for a given concept. It is equivalent to a logical expression where concepts are connected by the specified semantic operator (for example the boolean **AND** or the quantification operator **AT MOST**). The And meta-concept <bumpy:0, cracked:0, disordered:0, interlaced:0, lined:1, marbled:1, netlike:0, smeared:0, spotted:0, uniform:1, whirly:0 >is equivalent to the logical expression: $\exists$x s.t. (x = lined) $\wedge$ (x = marbled) $\wedge$ (x = uniform). A meta-concept description can therefore be soundly linked to a logical formula. Moreover, it is interpreted as: the texture distribution consists of lined, marbled and uniform textures. Therefore, the semantic interpretation of a meta-concept structure is a natural language sentence. We are then able to associate a user-formulated query through natural language interaction with a meta-concept and vice versa (e.g., a meta-concept characterizing an image index description is linked to a natural language sentence).

In our framework, each IO is linked to a meta-concept (either index or query) in the object facet and color and texture subfacets.

4.1 Object-based characterization

Each IO is linked to an *object meta-concept* oc $\in C_{\text{obj}}$, supported by a vector structure **o** with Card_{obj} elements corresponding to the highlighted object concepts. Values $o[i]$, $i \in [1,\text{Card}_{\text{obj}}]$ are booleans stressing that the considered IO is characterized by the object concept $c_{\text{obj}}[i]$.

In Fig. 3, the object meta-concept <sky:0, ground:0, field:0, …, huts:1, people:0…> characterizes Io2.

4.2 Texture characterization

Each IO is linked to a conceptual structure characterizing its texture distribution. The importance of highlighting symbolic texture features is correlated to the processing of user queries without making use of low-level visual features.

In order to integrate texture characterization within a symbolic multimedia (image) information retrieval framework, we moreover specify conceptual structures correlated to several types of user queries. Our approach is based on taking into account a language consisting of three boolean operators. A user shall be able to associate object concepts with a conjunction of texture categories such as in Q1: "*bumpy* and *cracked* roads", a disjunction of texture categories such as in Q2: "*brick-like* or *cracked* floors" and a negation of texture categories such as in Q3: "trees with non-*interlaced* leaves".

We introduce the conceptual structures correlated to these query types which characterize the symbolic texture distribution of IOs. We distinguish texture index meta-concepts which feature the texture distribution of IOs belonging to image index documents from query texture meta-concepts which translate texture distributions specified within queries. The latter extend the basic texture categories by taking into account the boolean semantic operator expressed.

4.2.1 Index structures

Each IO is indexed by a texture index meta-concept (TIC $\in C_{\text{ind\_tx}} \subset C_{\text{tx}}$). A TIC is supported by a vector structure **T** with 11 elements corresponding to texture categories t_i. Values $\mathbf{T}[i]$, $i \in [1, 11]$ are booleans stressing that the texture distribution of the considered IO is characterized by the texture category t_i. For example, in Fig. 3, the TIC <B:0,C:0,D:0,I:0,L:1,M:0,N:0,S:1,Sp:0,U:0,W:0>

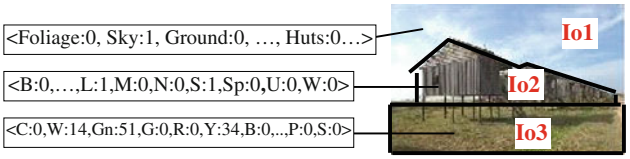

Fig. 3 Partial index representation of an example image

characterizes Io2 and can be interpreted as the texture distribution consists of lined and smeared textures.

4.2.2 *Query structures*

Three types of texture query meta-concepts (TQC $\in C_{q\_tx} \subset C_{tx}$) are specified to support the previously defined query types. And texture meta-concepts (ATCs) represent the signal distribution of an IO by a conjunction of texture categories; Or texture meta-concepts (OTCs) by a disjunction of texture categories and Not texture meta-concepts (NTCs) by a negation of texture categories. They are, respectively, characterized by vector structures $\mathbf{T}_{and}$, $\mathbf{T}_{or}$ and $\mathbf{T}_{not}$ such that values $\mathbf{T}_{and}[i]$, $\mathbf{T}_{or}[i]$ and $\mathbf{T}_{not}[i]$, $i \in [1, 11]$ are booleans stressing that the texture category t_i is an element of the conjunction, disjunction and negation of texture categories mentioned in the query. The ATC <B:1, C:1…I:0,L:0,N:0…>**and**, the OTC <B:0,C:1…I:0,L:0, N:1…>**or** and the NTC <B:0,C:0…I:1,L:0,N:0…>**not**, respectively, correspond to the texture distributions expressed in queries Q1, Q2 and Q3. The correspondence between TICs and TQCs is achieved through partially ordered lattices which are organized, respectively, to the type of the query processed.

4.3 Color characterization

In order to integrate color information within a symbolic model for multimedia information retrieval, we specify conceptual structures correlated to several types of queries formulated by the user. The approach developed in this paper is based on taking into account a rich query language consisting of six quantification operators representing the three categories of semantic quantification in [19]:

- A user is able to specify numerical quantifications linked to the proportion of pixels corresponding to the highlighted color categories through operators *at least* and *at most* such as Q4: "images with a cloudy sky (*At Most* 25% of cyan)" and Q5: "images with lake water (*At Least* 25% of grey)".

We also specify queries with literal quantifiers (*Mostly*, Few) easier to handle for a non-expert user and therefore less interested in a precise quantification of the highlighted color categories. These queries are such as Q6: "vegetation *mostly* green" and Q7: "flowers with *few* red".

We finally introduce the comparative quantifiers allowing comparing distributions of color categories within IOs. They are used in queries such as Q8: "sky with *more* cyan than grey".

We introduce the conceptual structures characterizing the symbolic color distribution of IOs and correlated to the type of queries defined. These structures extend the basic concepts by taking into account the quantification operator expressed. We distinguish in our approach the color index meta-concepts (CIC $\in C_{ind\_col} \subset C_{col}$) characterizing the color distribution of IOs within index images (thus carrying a conjunctive semantic) from the color query meta-concepts (CQC $\in C_{q\_col} \subset C_{col}$) which themselves translate the color distributions within queries. The latter can convey a semantic not limited to a trivial conjunction and which corresponds to the specification of several quantification operators for query formulation. We will therefore specify a formal framework for establishing a link between color index and query concepts.

4.3.1 *Index structures*

Each IO is indexed by a color index meta-concept (*CIC*) which features its color distribution by a conjunction of color categories and their corresponding integer pixel percentages. It is supported by a vector structure C with values $C[i]$ providing the pixel percentage of color category c_i. For example, in Fig. 3, the CIC <C:0,W:14, Gn:51,G:0,R:0,Y:34,B:0,…, P:0,S:0> characterizes Io3 and can be interpreted as the color distribution consists of 14% of white, 51% of green and 34% of yellow.

4.3.2 *Query structures*

- Our conceptual architecture is powerful enough to handle an expressive query language integrating object concepts and color characterization through quantified operators. We specify three categories of color query meta-concepts linked to the semantic of the operators defined for query color characterization:
- Numerical color meta-concepts linked to queries Q4 and Q5 which involve operators *At Most* and *At Least*, respectively. *At Most* color meta-concepts (*AMCCs*) and *At Least* color meta-concepts (*ALCCs*) represent the color distribution of an IO by a conjunction of color categories and, respectively, their associated maximum (translating the keyword *At Most* in a query) and minimum (translating the keyword *At Least*) pixel percentage. They are, respectively, supported by vector structures C_{ap}, C_{am} such that values $C_{ap}[i]$ and $C_{am}[i]$, $i \in [1, 11]$ are percentages (integer) stressing that the color category c_i is an element of the conjunction, disjunction and negation of texture categories mentioned

in the query. For instance, the AMCC <c:25,w:0,gn:0,g:0,r:0…>$_{\mathbf{AM}}$ and the ALCC <c:0,w:0,gn:0,g:25,r:0…>$_{\mathbf{AL}}$, respectively, correspond to the signal distributions expressed in queries Q4 and Q5.

- Literal color meta-concepts associated to queries Q6 and R7 making, respectively, use of operators *Mostly* and *Few*. *Mostly* and *Few* color meta-concepts represent the color distribution of an IO by a conjunction of color categories whose pixel proportions are, respectively, majoritary (translating the keyword *Mostly* in a query) and minoritary (translating the keyword *Few*). They are, respectively, supported by vector structures C_{maj} (C_{min}) such that values $C_{\text{maj}}[i]$ ($C_{\text{min}}[i]$), $i \in [1, 11]$ are booleans translating that color catgory c_i is linked to a majoritary (minoritary) pixel proportion within the color distribution of the considered IO. For example, the *Mostly* color meta-concept <c:0,w:0,gn:1,g:0,r:0…>$_{\mathbf{Mostly}}$ and the *Few* color meta-concept <c:0,w:0,gn:0,g:0,r:1…>$_{\text{Few}}$, respectively, correspond to the color distributions expressed in queries Q6 and Q7.
- Literal color meta-concepts associated to queries Q6 and R7 making, respectively, use of operators *Mostly* and *Few*. *Mostly* and *Few* color meta-concepts represent the color distribution of an IO by a conjunction of color categories whose pixel proportions are, respectively, majoritary (translating the keyword *Mostly* in a query) and minoritary (translating the keyword *Few*). They are, respectively, supported by vector structures C_{maj} (C_{min}) such that values $C_{\text{maj}}[i]$ ($C_{\text{min}}[i]$), $i \in [1, 11]$ are booleans translating that color catgory c_i is linked to a majoritary (minoritary) pixel proportion within the color distribution of the considered IO. For example, the *Mostly* color meta-concept <c:0,w:0,gn:1,g:0,r:0…>$_{\mathbf{Mostly}}$ and the *Few* color meta-concept <c:0,w:0,gn:0,g:0,r:1…>$_{\text{Few}}$, respectively, correspond to the color distributions expressed in queries Q6 and Q7.
- Comparative color meta-concepts linked to the query Q8 involving the operators *More than* and *Less than*. *More/Less* color meta-concepts, correlated to two sets d_{more} and d_{less}, are linked to the color distribution of an IO whose pixel proportions of some color categories (elements of the set d_{more}) are more important than that of other color categories (elements of the set d_{less}). Elements of the sets d_{more} and d_{less} are, respectively, preceded by keywords *More* and *Less*. *More/Less* color concepts are characterized by a pair of vectors (C_m, C_l), each vector with a number of elements equal to the number of color categories. Values $C_m[i]$ and $C_l[j]$, $i,j \in [1, 11]$ are boolean values which are true when the proportions of the dominant color categories c_i are higher than those of the dominant color categories c_j. For example, the *More/Less* color concept (<c:1,w:0,gn:0,g:0,r:0…>$_m$, <c:0,w:0,gn:0,g:1,r:0…>$_l$) corresponds to the color distribution in query Q8.

Color meta-concepts are elements of partially ordered lattices which are organized, respectively, to the type of the query processed.

4.4 Index representation

To build a meta-conceptual image index representation I, object meta-concepts, TICs and CICs are automatically derived for each IO. We provide in Fig. 3 a partial index representation for an example image with the object meta-concept related to Io1, the TIC related to Io2 and the CIC related to Io3.

5 The query module

5.1 Query expression

Our conceptual architecture is based on a unified full-text framework allowing a user to query over the visual semantics and signal information. This obviously optimizes user interaction since the user is in 'charge' of the query process by making his information needs explicit to the system.

To build a meta-conceptual image query representation Q, object meta-concepts, TQCs and CQCs are automatically derived for each IO. Without going into details, a simple grammar composed of a list of the previously introduced concepts is used to parse the user full-text query.

5.2 The matching process

The matching framework is based on Van Rijsbergen's logical model which measures to which extent the image document I satisifies the query Q through the exhaustivity function P:

$$\text{Relevance}(I, Q) = P(I \rightarrow Q). \quad (2)$$

P consists of two operations. It first checks that all concepts described within the query are also elements of the index representation. For this, we use lattice projection to compare query and index concepts. Then, for each selected image, we provide an estimation of its relevance with respect to the query, which corresponds to the quantitative evaluation of their similarity. It is given by the exhaustivity value between query q and index representation i:

$$\begin{aligned}\mathrm{EV}(q,i) = \mathrm{MAX}[&\Sigma_{\mathrm{OC}q\,\text{concept of}\,q,\,\mathrm{OC}_i\,\text{matching concept of}\,i}\,\mathrm{Imp}(\mathrm{OC}_i)\\ &+\mathrm{Cpt\_Match}\left(\mathrm{OC}_q,\mathrm{OC}_i\right)\\ &\times\Sigma_{\mathrm{TQC}q\,\text{concept of}\,q,\,\mathrm{TIC}_i\,\text{matching concept of}\,i}\\ &\times\mathrm{Cpt\_Match}\left(\mathrm{TQC}_q,\mathrm{TIC}_i\right)\\ &+\Sigma_{\mathrm{CQC}q\,\text{concept of}\,q,\,\mathrm{CIC}_i\,\text{matching concept of}\,i}\\ &\times\mathrm{Cpt\_Match}\left(\mathrm{CQC}_q,\mathrm{CIC}_i\right)]\end{aligned}$$

The Imp function measures the 'importance' of an object concept within an image. It is both proportional to the size of the corresponding visual object and its global localization with respect to the image center. The *Cpt_Match* function is the negative Kullback–Leibler divergence between the probabilities of visual semantic, texture and color query concepts which are themselves certain (i.e. $P(\mathrm{OC}_q)$, $P(\mathrm{TQC}_q)$ and $P(\mathrm{CQC}_q) = 1$) and:

- the posterior recognition probabilities of object concepts of graph i
- the average of posterior recognition probabilities of texture and color categories within matching index texture and color concepts of graph i.

5.3 Organization of the lattice of object concepts

The conceptual structures are organized within latticed-based structures defined by a specific/generic partial order (corresponds to a specialization operation in Fig. 4). To derive the lattice of object concepts, several experimental studies presented in [2] have led to the specification of 20 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 object concepts to be learnt and automatically extracted are specified. These are further enriched with the concepts of the LSCOM-lite taxonomy (which includes concepts related to the characterization of multimedia news information and in particular individuals' identities in the TRECVid topic search task) and a part of the object lattice is provided in Fig. 4.

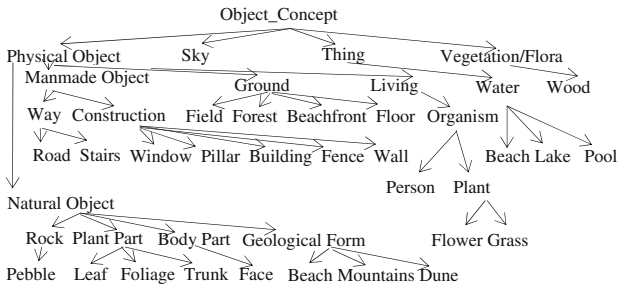

Fig. 4 A part of the lattice of object concepts

6 Validation experiments

The SIR (Signal/object/semantic integration for Image Retrieval) prototype implements the theoretical framework and validation experiments are carried out on both a corpus of consumer photographs and the TRECVid corpus of image keyframes extracted from news videos. We choose to deal with these multimedia collections instead of the Corel professional collection since it has been argued and demonstrated experimentally that the latter is much easier to annotate and retrieve; and in fact does not capture the difficulties inherent in datasets used in real world. We first introduce these empirical test collections then deal with the experimental Object Level characterizations, detailing the algorithmic processes for the object highlighting framework and its evaluation. We finally evaluate the Semantic Level characterization with queries involving multiple characterizations of the visual content.

6.1 Test collections

The first collection consists of 2,500 heterogeneous consumer photos. The images are of resolution 256 × 384, in both portrait and landscape layouts and comprise outdoor and indoor scenes. For outdoor images, the content varies from natural landscape (beach, lakeside, river, pond, park, forest, garden, mountain, rocky area…) to city scenes (urban area, rural area, crowded street, market, road with vehicles, swimming pool, temple, mosque, castle…) from different countries and cultures (Singapore, France, Belgium, China, Cambodia, Malaysia, Indonesia…). The indoor images are taken with different focuses (portrait of single person or a few people, groups of different sizes, people eating, cultural performance, wedding ceremony, interior layout, display of objects like painting, toys, antique collection…). In both outdoor and indoor images, the subject of focus could be people (or faces in photo frame), statues, animals, flowers, buildings (or their miniature in theme park), etc., and their mixture with occlusion, taken with different postures, during the day or at night, from different viewpoints, and at different distances. We can also find photographs of low quality which are however kept for processing in order to illustrate the heterogeneity of this particular type of image collections. Figure 5 illustrates some of the photos of poor quality (e.g., faded, overexposed, blurred, etc.).

Unlike professional images, which are well defined, with sharp contrast and homogeneous signal distributions;

Fig. 5 Photographs of low quality kept in the collection

or domain-specific images such as medical images, which have a clear classification and are usually attached with semantic annotation, consumer image content varies significantly. Image segmentation is furthermore challenging due to the heterogeneity of color and texture regions. Also, the lack of content annotation does not make it possible to assume availability of text for straightforward joint visual/text characterization.

The TRECVid_04 corpus consists of 70 h of US broadcast news video in MPEG-1 format. It comprises 128 videos segmented in 33,366 shots, each one itself represented by 1 or more keyframes (cf. Fig. 8 for keyframe samples). A total of 48,817 keyframes are specified. The key-frame extraction process is here integrated with the processes of shot segmentation. Each time a new shot is identified, the keyframe extraction process is invoked, using parameters already computed during shot boundary detection. These parameters are related to visual data such as color or camera motion descriptors. Keyframe selection differs depending on the application needs. For example, we require only a few keyframes (1–2) for a video-captured meeting since camera motions are sparse. However, in broadcast news documents, we find more animation (related to visual aspects) which entails highlighting more shots and keyframes. In our approach, the keyframes are selected as the most stable images of a given shot and there are up to 10 keyframes representing a shot.

6.2 Experimental Object Level characterization

As far as the feature extraction processes are concerned, our algorithm is summarized below:

- Given an image in the index corpus.
- We apply a rectangular grid to it highlighting the nb_grid rectangular regions of size 35 × 35 pixels which override of 12 pixels in [Ox) and [Oy).
- For each rectangular region, we characterize the color and texture categories as presented in Sects. 2.1 and 2.2 respectively. It is then described by a 22-dimension structure.
  - After a step of low-level characterization in the RGB space, the color categories with their probabilities are highlighted in the perceptual HVC space.
  - For texture characterization, the 49-dimension Gabor energy vectors are linked to the texture categories.

For the learning process, 554 rectangular image regions are cropped from 138 images taken from the collection of 2,500 home photographs. Additionally, 885 keyframe regions are extracted from the TRECVid_03 collection. Samples are provided in Fig. 6. 965 (i.e., two-thirds) of them are used as training data and the remaining one-third (i.e., 474) as test data for generalization performance. In other words, both the training and test data utilize only a small percentage of the original collections. Color and texture features are computed for each training region as an input vector for the neural network.

The learning algorithm is summarized as follows:

- Given a 'positive' visual object, i.e., corresponding to the object concept being learned:
  - We apply to it a rectangular grid highlighting the rectangular regions with size 35 × 35 pixels
  - We extract the color and texture categories for each rectangular region with their associated recognition probabilities.
  - Features corresponding to the block are used as training input for the probabilistic visual semantic tagging framework.

For the recognition step, the algorithmic process is based on five steps:

- Given a physical keyframe segmented into rectangular regions.
- For each object concept, we use the probabilistic classifier which provides an output value.
- We obtain for each rectangular region the probabilities linked to the object concepts.

Fig. 6 Example of cropped image regions used as training data (*top down*, *left* to *right*): face, people, crowd, body part, sky (clear), sky (cloudy), sky (*blue*), floor, beachfront, field, pool, pond, lake, foliage, flower, trunk, mountain, pebble, pillar, building, stairs, wall (*white*), wall (wooden), fur, cycle, fire

- We consider as the representative object concept the one associated to the maximum recognition probability.
  - In case of a "conflict" (i.e. a rectangular region such as two distinct object concepts have the same maximum probability values), the decision will be based on taking into account the object concepts with the maximum recognition probabilities in adjacent regions.
- We agglomerate the blocks with respect to the protocol described in Sect. 3.2 so as to highlight the image objects.

6.3 Evaluation of the Semantic Level characterization

As opposed to trivial object concept queries, we wish to retrieve images that represent elaborate image scenes through queries involving multiple characterizations at the Semantic Level (as proposed in the TREC multimedia track). For this evaluation, we use two validation corpora: the first is the previously introduced collection of real-world photographs and the second is the TRECVid_04 keyframe corpus.

6.3.1 Evaluation on the collection of home photographs

For this, we specify 45 queries implying non-trivial information needs based on object concepts with additional signal characterizations such as *swimming-pool water* or *interlaced foliage* with their ground truths.

For each proposed query, we construct relevant textual query characterizations at the Semantic Level using corresponding object and signal concepts as input to SIR (e.g. 'water mostly cyan' for *swimming-pool water* or 'people lined' for *lined people*). S_1 processes 3 series of 3 random relevant photographs for each query (they correspond to swimming-pool water, lined people in our example queries). Also these queries are translated in relevant textual symbolic entities to be processed by the semantic framework of S_2 ('Find images with water' for *swimming-pool water,* 'Find images with people' for *lined people*). Then to refine the results, 3 random relevant photographs are selected as input to the RF framework.

Recall/precision curves of Fig. 7 illustrate the average results obtained for queries involving object concepts and signal characterizations: the curve associated with the *SIR* legend illustrates the results in recall and precision obtained by SIR, 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 SIR (0.4292) is approximately 78.54% higher over the average precision of the VK system (0.2404) and approximately 35.61% higher over the

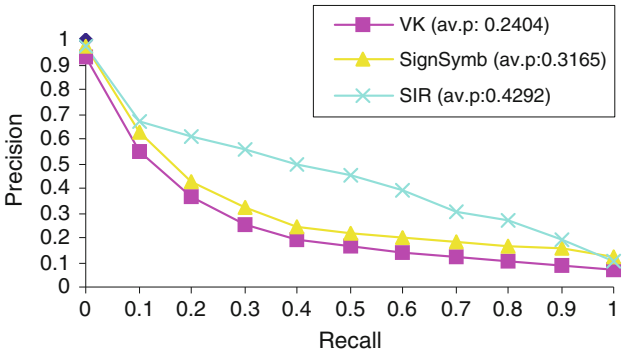

Fig. 7 Recall/precision curves for the signal/object queries

average precision of the loosely coupled state-of-the-art system (0.3165). We notice that improvements of the precision values are significant at all recall values.

6.3.2 Evaluation on the TRECVid corpus

First, image objects within the 48,817 image keyframes are automatically assigned an object concept as presented in Sect. 3 and characterized with conceptual object, color and texture structures presented in Sects. 4.1, 4.2 and 4.3.

The search task is based on *topic retrieval* where a topic is defined as a formatted description of an information need, therefore involving multiple characterizations such as images, audio, text… The complexity inherent in topic search revolves around the difficulty to design the intended meaning and interrelationships between the various characterizations. We therefore design the evaluation task in the context of manual search, where a human expert in the search system interface is able to interpret a topic and propose an optimal query to be processed by the system. 15 multimedia topics provided in Table 1 and their ground truths developed by NIST for the search task express the need for video concerning people, things, events, locations and combinations of the former. Topics are designed to reflect many of the various sorts of queries users propose: requests for documents with specific people or people types, objects or instances of object types, activities or locations or instances of activity or location types.

We propose in Table 1 their formulation at the Semantic Level of our architecture as a SIR query. The latter is compared with signal-based and object-based systems operating manual search on visual features. The *signal-based system*, representative of the category of systems tackling the Signal Level, is based on a query-specific combination of visual content-based retrieval: color characterization consists of the concatenation of a global 166-dimensional HSV color correlogram and 3 × 3-grid based 81-dimensional Lab color moments; texture characterization is based on the concatenation of a global 96-dimensional

Table 1 TRECVid topics, their query transcription and equivalent logical formulations at the Semantic Level

TRECVid topic	SIR query transcription
Street scene with multiple pedestrians in motion and multiple vehicles in motion	People and cars
One or more buildings with flood waters around it/them	Buildings and water smeared
One or more people and one or more dogs walking together	People and fur
US Capitol dome	IO mostly grey and lined
Hockey rink with at least one of the nets fully visible from some point of view	IO netlike and IO mostly white and marbled
Person hitting a golf ball that then goes into the hole	People and IO mostly green and uniform
One or people going up or down some visible steps or stairs	People and stairs
Handheld weapon firing	IO black and fire
One or more bicycles rolling along	Bicycles lined
Tennis player contacting the ball with his or her tennis racket	People and IO netlike and mostly yellow
Bill Clinton speaking with at least part of a US flag visible behind him	Face and IO lined and blue and white and red
One or more horses in motion	Fur
One or more skiers skiing a slalom course with at least one gate pole visible	People and ground mostly white
One or more buildings on fire, with flames and smoke visible	Buildings and fire
One or more signs or banners carried by people at a march or protest	Crowd and IO white

co-occurrence feature and a 3 × 3 grid-based 27-dimensional Tamura feature. Each query is manually formulated as a boolean or a weighted average combination of queries based on visual examples [20]. The *object-based system*, representative of the category of systems tackling the Image Object Level, relies on a generative probabilistic model (Gaussian Mixture Model). Queries are created by manual construction and selection of visual examples [21].

We propose the top retrieval results for four multimedia topics in Fig. 8. Also, Table 2 details the precision at n documents for the compared systems. We can note that the precision at 10 documents for SIR (0.113) is approximately 116.48% higher than this of the semantic-based system (0.052) and more than 25 times better than this of the signal-based system (0.004). This clearly indicates that on average the first keyframe images returned by SIR are particularly relevant compared to the first keyframe images retrieved by other systems. SIR is therefore precision-oriented, which is an interesting property since according to studies related to users' behavior, an

Fig. 8 Top 4 retrieval results for topics 127 (one or more people and one or more dogs walking together), 130 (hockey rink with at least one of the nets fully visible from some point of view), 136 (person hitting a golf ball that then goes into the hole), 140 (one or more bicycles rolling along)

Table 2 Precision at *n* documents for the compared systems

Precision	SIR	Semantic-based	Signal-based
At 5 docs	0.1652	0.0609	0.0087
At 10 docs	0.1130	0.0522	0.0043
At 15 docs	0.0870	0.0493	0.0116
At 20 docs	0.0891	0.0391	0.0087
At 30 docs	0.0754	0.0319	0.0101
At 100 docs	0.0452	0.0213	0.0104
At 200 docs	0.0322	0.0159	0.0096
At 500 docs	0.0165	0.0118	0.0074
At 1,000 docs	0.0105	0.0092	0.007

individual is mostly interested in the first retrieval results provided by a system.

In the framework of the TRECVid experiments, we notice that the results obtained by the compared systems are much poorer than those obtained considering the corpus of home photographs, which is explained by the very low image quality of keyframes extracted from videos. This strongly penalizes all extraction processes operating at the image level (e.g., color, texture…).

7 Conclusion and perspective

To address the difficulty to reconcile the heterogeneous ways of interpreting the multimedia visual content (and in particular the gap between the signal and semantic interpretations), we have proposed an architecture based on the use of multiple representations of the visual content corresponding to three abstraction levels, namely the Signal, Object and Semantic levels.

At the Signal Level, we specify a framework mapping the extracted low-level features to high-level visual information through the use of a population of learning agents which have the ability to perceive signals (color or texture), categorize their perception and lexicalize their symbolic representation.

At the Object Level, we highlight a correspondence between the signal information and object concepts (such as *mountains*, *sky*, *grass*…). The automatic object-based indexing framework, based on a statistical model which considers the joint distribution of semantic concepts and symbolic signal information, addresses the curse of dimensionality contrary to traditional frameworks considering high-dimensional spaces of low-level extracted signal features.

At the Semantic Level, signal and object characterizations are coupled within a logic-based framework. The latter is instantiated by a knowledge representation formalism allowing to define an expressive query language consisting of several operators as well as a theoretically sound matching module for query processing.

Empirically, the SIR prototype implements the theoretical proposition and validation experiments are carried out on both a corpus of consumer photographs and the TRECVid corpus of image keyframes instead of "easy-to-process" professional collections. They allow us to stress the gain in precision at both Object and Semantic levels of characterization of our framework in comparison with state-of-the-art architectures.

There are however limitations when learning semantic concepts through the use of low-level features. In such frameworks, solutions to multimedia indexing and retrieval could only be applied using broad semantic concept detectors, e.g. "sky" or "foliage", therefore leading to restrained index vocabularies. Perspectives in large-scale multimedia indexing and retrieval would consist in using the Web and in particular the widely available contextual image information (however, not only restricted to tags available in social websites). However, it is believed that indexing based on contextual image information solely gives tolerable results. But if used jointly with both the signal features and semantic concepts derived from the visual content, the precision of image search engines could be further improved. This idea will be further explored in subsequent works.

References

1. Smeulders, A., et al.: Content-based image retrieval at the end of the early years. IEEE PAMI **22**(12), 1349–1380 (2000)
2. Mojsilovic, A., Rogowitz, B.: Capturing image semantics with low-level descriptors. ICIP, pp.18–21 (2001)
3. Blei, D.M., Jordan, M.I.: Modeling annotated data. SIGIR, pp. 127–134 (2003)
4. Carneiro, G., Chan, A.B., Moreno, P.J., Vasconcelos, N.: Supervised learning of semantic classes for image annotation and retrieval. IEEE Trans. Pattern. Anal. Mach. Intell. **29**(3), 394–410 (2007)
5. Feng, S., Manmatha, R., Lavrenko, V.: Multiple Bernoulli relevance models for image and video annotation. CVPR **2**, 1002–1009 (2004)
6. Jeon, J., Lavrenko, V., Manmatha, R.: Automatic image annotation and retrieval using cross-media relevance models. SIGIR, pp. 119–126 (2003)
7. Li, J., Wang, J.Z.: Real-time computerized annotation of pictures. IEEE PAMI **30**(6), 985–1002 (2008)
8. Liu, J., et al.: Dual cross-media relevance model for image annotation. ACM MM, pp. 605–614 (2007)
9. Jin, Y., et al.: Image annotations by combining multiple evidence and wordNet. ACM MM, pp. 706–715 (2005)
10. Srikanth, M. et al.: Exploiting Ontologies for Automatic Image Annotation. ACM SIGIR, pp. 1349–1380 (2005)
11. Bradshaw, B.: Semantic based image retrieval: a probabilistic approach. ACM MM, pp. 167–176 (2000)
12. Lim, J., Jin, J.S.: A structured learning framework for content-based image indexing and visual query. Multimed. Syst. **10**(4), 317–331 (2005)

13. Town, C.P., Sinclair, D.: CBIR Using Semantic Visual Categories. TR2000-14, AT&T Labs Cambridge (2000)
14. Mulhem, P., et al.: Advances in Digital Home Image Albums. Multimedia Systems and Content-Based Image Retrieval, Idea Publishing, chapter IX, pp. 201–226 (2003)
15. Mechkour, M.: EMIR^2: An Extended Model for Image Representation and Retrieval. DEXA, pp. 395–404 (1995)
16. Meghini, C., et al.: A model of multimedia information retrieval. J. ACM **48**(5), 909–970 (2001)
17. Berlin, B., Kay, P.: Basic Color Terms. Their Universality and Evolution. UC Press, Berkeley (1991)
18. Bhushan, N., et al.: The texture lexicon: understanding the categorization of visual texture terms and their relationship to texture images. Cogn. Sci. **21**(2), 219–246 (1997)
19. Peters, S., Westerthal, D.: Quantifiers. MIT Press, Cambridge, MA (2002)
20. Kender, J.R., et al.: IBM Research TRECVID Video Retrieval System. In: Online Proceedings of the TREC Video Retrieval Evaluation. http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html#2004
21. Ianeva, T., et al.: Probabilistic approaches to video retrieval. TREC video retrieval evaluation online proceedings. http://www-nlpir.nist.gov/projects/tvpubs/tvpapers04/cwi-twente.pdf
22. Gong, Y., et al.: Image indexing and retrieval based on color histograms. Multimed. Tools App. **II**, 133–156 (1996)