Symbolic Photograph Content-Based Retrieval

Philippe Mulhem
IPAL CNRS
21 Heng Mui Keng Terrace
Singapore 119613
(+65) 6874 8212
mulhem@lit.a-star.edu.sg

Joo Hwee Lim
Laboratories for Information Technology
21 Heng Mui Keng Terrace
Singapore 119613
(+65) 6874 6671
joohwee@lit.a-star.edu.sg

ABSTRACT

Photograph retrieval systems face the difficulty to deal with the different ways to apprehend the content of images. We consider and demonstrate here the use of multiple index representations of photographs to achieve effective retrieval. The use of multiple indexes allows integration of the complementary strengths of different indexing and retrieval models. The proposed representation supports multiple labels for regions and attributes, and handles inferences and relationships. We define links between indexing levels and the related query modes. The experiment conducted on 2400 home photographs shows the behavior of the multiple indexing levels during retrieval.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *query formulation, retrieval models, search process.*

General Terms

Algorithms, Experimentation, Theory.

Keywords

Image Access, Fusion/Combination.

1. INTRODUCTION

Content-base image retrieval systems have been the subject of many research works in the 90s. They allowed discovering adequate processes related to image indexing, efficient content representation for retrieval according mainly to low level features like colors/textures/shapes, and query modes related to the specificity of the image data. At the same time, different query input modes, inspired from text information retrieval, have been used for such image retrieval systems.

The systems that are dedicated to image retrieval (as described in [25]) use mainly features like colors, textures and shapes as bases for the representation of image content. For applications dedicated to non-experts (as described in [10]), users prefer to use the elements present in images than the rough color/textures/shapes. The work of the MPEG-7 committee [18] also represents such elements for video documents. Inspired by previous work like EMIR2 [17] and VIMSYS [9], we represent three facets for the representation of photographs content:

- The first level, namely *Feature Level*, represents numerical abstractions of image regions. Such abstractions express the colors, textures and shapes of visible elements in photographs. Systems like Virage[1], NETRA[16], QBIC[6, 7], VisualSEEk[26, 27], or proposed by [22], consider this level of representation. Such approaches do not at all rely on symbolic representation of images, however they have the great advantage of being fully automatic and to be usable in many contexts with (almost) no tailoring. In the remaining of the paper, this level one is not considered because they are relatively further from the semantic level of human users' query expectation.
- The second level (namely "*Visual Object Level*") of representation supports the notion of labeling of photograph parts. This level intends to bridge a gap between the signal aspects (first level) and the symbols that represent the image's content. If the process of labeling is manual, then the labels may be certain. In the context of automatic labeling processes [4, 14, 30], the labeling may be uncertain, leading to multiple potential labels for image regions. This level also supports the representation of the elements of the images.
- The third level (namely "*Relational Concept Level*") is dedicated to represent relations between image elements, or the explicit characteristics of the images elements of the second level. The characteristics of the image elements may be of several natures: absolute spatial (like the position of an element in the photograph), descriptive (like face expressions or person postures). The relations between image elements may be relative spatial (like relative positions between elements) or actions (like the fact that "a person is carrying a hat" for instance). Thus this level is able to weigh the relationships and concepts to reflect their importance in the image representation. The third level representation may be able to support inference (as mentioned in [20]), for instance IsA hierarchies of labels and relations, and to include properties of relations like symmetry, in a way to extend the retrieval capabilities.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
CIKM'02, November 4-9, 2002, McLean, Virginia, USA.
Copyright 2002 ACM 1-58113-492-4/02/0011…$5.00.

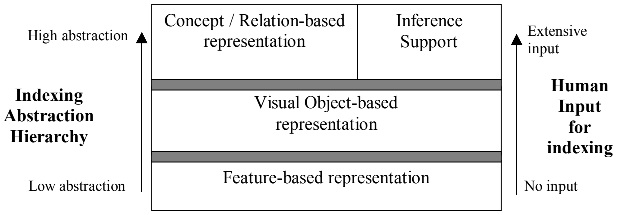
High abstraction
Low abstraction
Indexing Abstraction Hierarchy
Concept / Relation-based representation
Inference Support
Visual Object-based representation
Feature-based representation
Extensive input
Human Input for indexing
No input

Figure1. Indexing levels.

If we examine a notion of abstraction targeted at each level (left part of Figure 1), the Feature Level corresponds to low abstraction (because it remains at a signal level), the Visual Object Level is a medium abstraction representation (because it abstracts the signal but does not go further), the Relational Concept Level is a high abstraction level because it assumes a priori relationships between relations and concepts and allow interpretation of scenes. The structure of Figure 1 shows that the three levels of indexes are obviously not independent from each other. However, the gray parts that represent the transition between the levels are far from being easy tasks: going from signal features to visual objects is usually based on some learning process or on manual image annotations, whereas going from Visual Object Level to Concepts/Relations needs human input. The right part of Figure 1 exhibits this fact by showing that human input increases when the abstraction representation of the image content becomes higher. This human input during the indexing process is the way to give relevant human knowledge to the indexing system, keeping in mind that the retrieval will take advantage of this input. In this paper, we propose a way to tackle these different levels as well as steps to bridge the gaps among the three levels.

Another very important aspect of document and image retrieval is related to the user interactions supported by the retrieval system. According to previous works, the user can specify a query for image retrieval by different query modes:

- *Query By Example* (QBE). This kind of query was proposed in Virage [1] and QBIC [6] in the early 90s.The query by example (related to the well known Relevance Feedback technique of Information Retrieval) exploits a user's input in the form of one or several images and generates a query. It suffers from the bootstrapping problem. Moreover, the user's need is implicit: the system has to guess a query using its knowledge of the given image(s) and of the user. So, the query input mode is very simple for the user (in terms of actions [19]), but the implicit information is very difficult to extract for the image retrieval system if we consider uncertain labels at the level 2 or the complex representations of level 3. Because of this problem, the use of low-level abstraction representations in conjunction with QBE modes is common. A way to deal with the abstraction problem with QBE is to provide "penetrable query by example" (inspired from [11]). The system is supposed to provide more accurate results, but the user has to support this additional step to reduce the ambiguities.
- *Query By Canvas* (QBC). The QBIC, Virage and VisualSEEk systems propose this kind of interface. A user draws an example of what he/she is looking for. This drawing is considered as a schematic representative of the ideal searched images. In this case, the user has to make extra efforts to "fit" into the system representation, so it is more difficult to the user to express his/her needs at the articulation level [19]. When considering simple interaction (i.e. the user selects the features first (e.g. color, texture, shapes) and then draws on the working space the regions with the selected features, representations of level 1 and level 2 may be used. This interaction lowers ambiguities because the drawing is supposed to focus on the relevant aspects of the searched images. However the user expects the system to understand his/her query semantics represented by the drawing that is based on primitive features (e.g. blue rectangle, red circle etc). Lim [13,14] elevates the semantics level of the drawing to layout of visual objects with spatial constraints with very promising retrieval results on family photographs. Nevertheless, the abstraction level supported by such query mode is higher than that allowed by QBE, even if ambiguities may arise from the interpretation of the relations among objects.
- *Query By Symbols* (QBS) [17, 20]. The query by symbols mode requires a user to explicit specify each of the elements, features and relations that he/she is looking for. On one hand, this expression is obviously more demanding for the user, but on the other hand the expression is closer to his/her actual information need. This query mode is able to handle highly abstracted representations of level 3.

Figure 2 summarizes the links among the three query modes described above and the abstraction level representations targeted by the corresponding query expressions. The three query modes have both their strengths and weaknesses, and the choice of using one or the other should be left to the user.

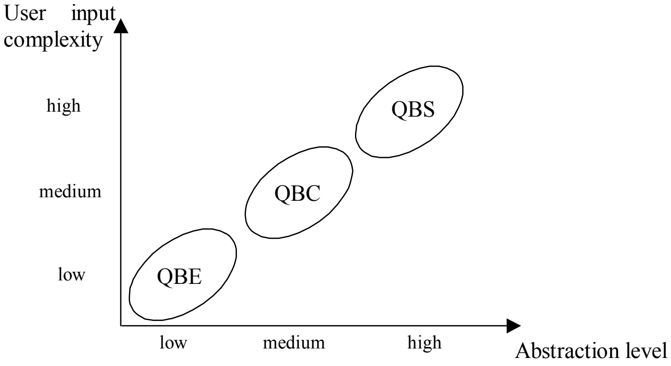

Figure 2. User input complexity vs. query abstraction levels.

In the following of this paper, we describe in section 2 the way we use two different approaches, namely Visual Keywords (VK) and Conceptual Graphs (CG), to support the indexing levels 2 and 3. Section 3 is dedicated to present the experiment we conducted on a set of 2400 home photographs, showing how the two approaches are integrated. We conclude in Section 4.

2. SYMBOLIC BASED INDEXING AND RETRIEVAL

2.1 Overview

As described above, a content-based image retrieval system should be able to support a wide range of query input modes, from QBE to QBS. The representation of image content should support and to use simple and complex query expressions effectively. To realize the image content representations of levels 2 and 3 (Figure 1) as described in the introduction, our approach adopts Visual Keyword (VK) [13-15] and Conceptual Graph [17, 20] that together handle:

1. Uncertainty recognition values of image elements (level 2)
2. Spatial characteristics of image elements (level 3)
3. Relations among image elements (level 3)
4. Inference support (level 3)

We integrate the Visual Keyword and Conceptual Graph based representations because they complement each other in the following ways:

- The VK handle multiple fuzzy labels for image regions efficiently (level 2)
- The VK approach manages spatial attributes of image elements (level 3)
- The CG approach handles relations among image elements (level 3)
- The CG approach incorporates inference capabilities for the descriptions (hierarchies of concepts and of relations), and the retrieval operator of the CG approach is a deduction process (level 3)
- The weights of image elements and relations are represented by the CG approach.

Figure 3 presents graphically the indexing levels realized by the VK and CG approaches with respect to the image indexing levels 2 and 3 of Figure 1.

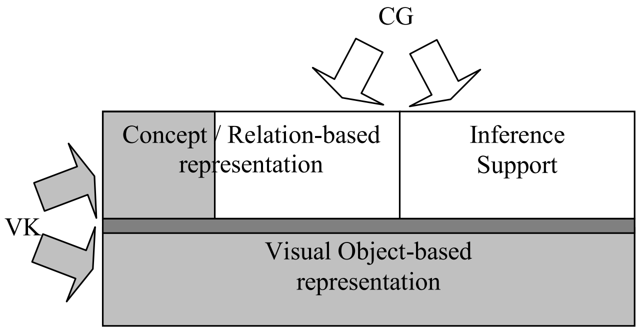
CG
Concept / Relation-based representation
Inference Support
VK
Visual Object-based representation

Figure 3. Indexing levels by VK and CG.

2.2 The Visual Keywords Approach

The Visual Keyword approach [13-15] is a new attempt to achieve content-based image indexing and retrieval beyond the feature-based (e.g. QBIC [7]) and region-based (e.g. VisualSEEk [26]) approaches. Visual keywords are intuitive and flexible visual prototypes extracted or learned from a visual content domain with relevant semantics labels. An image is indexed as a spatial distribution of visual keywords whose certainty values are computed via multi-scale view-based detection.

The indexing process has the following steps. First a visual vocabulary and thesaurus is constructed (keyword definition) from samples of a visual content domain. Then an image to be indexed is compared against the visual vocabulary to detect visual keywords (keyword detection) automatically. Last but not least, the fuzzy detection results are registered as a *Fuzzy Object Map* (FOM) and further aggregated spatially (spatial summary) into a *Spatial Aggregation Map* (SAM).

Visual keywords are visual prototypes specified or learned from domain-relevant regions of sample images. A set of labels S^{L}_{VK} is assigned to these visual prototypes as a vocabulary. Suitable visual features (e.g. color, texture) are computed for each training region and visual keyword into feature vectors. Figure 4 shows some examples of visual keywords used in our experiment.

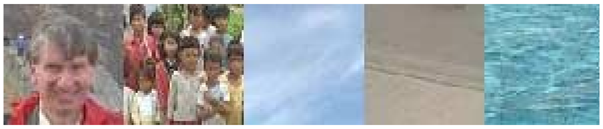

Figure 4. One visual keyword for the visual classes: face, crowd, sky, ground, water.

An image to be indexed is scanned with windows of different scales. Each scanned window is a visual token reduced to a feature vector to those of the visual keywords previously constructed. Figure 5 shows a schematic diagram of the architecture of image indexing (please refer to [14] pp.127-128 for details).

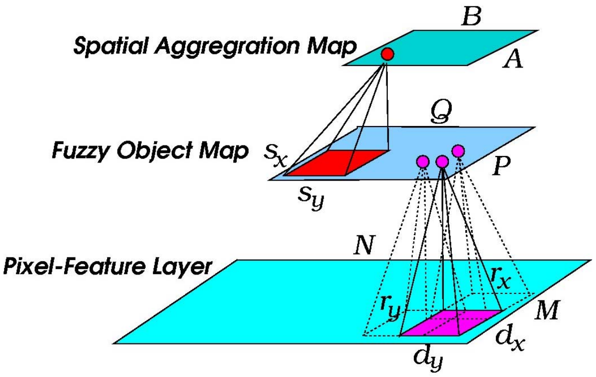

Figure 5. Automatic Indexing using VK.

Visual keywords can be regarded as co-ordinates that span a new pseudo-object feature space. The scale on each dimension is the fuzziness ($\in[0,1]$) of that visual keyword being detected at a specific spatial locality in the visual content.

As SAM summarizes the visual content of an image, similarity matching between two images can be computed as the weighted average of the similarities between the corresponding tessellated blocks of the images,

$$\lambda(x,y)=\frac{\sum_{(a,b)}\omega(a,b)\lambda(a,b)}{\sum_{(a,b)}\omega(a,b)}$$ and

$$\lambda(a,b) = 1 - \frac{1}{2}\left|SAM_x(a,b) - SAM_y(a,b)\right|$$

where $\omega(a,b)$ is the weight assigned to the block *(a,b)* in SAM and $\lambda(a,b)$ is computed using city block distance |.| between two corresponding blocks *(a,b)* of the images. For the query processing of QBE in our experiment, we adopted the tessellation (and their relative weights $\omega(a,b)$) as depicted in Figure 6.

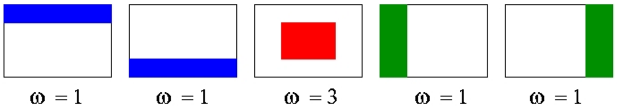

Figure 6. Tesselation and similarity matching in QBE.

When multiple query examples (say q_1 to q_k) are selected for QBE, the RSV for an image d in the database is computed as $RSV(SAM_d, \{SAM_{qi}\}) = max_i(\lambda(d,q_i))$.

In summary, the VK approach provides a simple, compact, and efficient representation for multiple labeling of image elements. The simple and fast similarity matching process also takes care of absolute spatial relations as specified in the tessellation of SAM.

2.3 The Conceptual Graphs based Approach

The chosen framework that handles concepts and hierarchies of concepts as well as relations and hierarchies of relations easily is the knowledge representation formalism called Conceptual Graphs [28, 29]. The formalism has already been used on photograph content representation [17, 20]. It has also been shown to be compatible with inverted file implementation [20]. Conceptual graphs are bipartite finite oriented graphs composed of concept nodes and of relation nodes. Concepts node are composed of a concept type and a referent (generic or individual). A generic referent, noted *, denotes the existence of a referent, while an individual referent (like #IMG0232 in figure 7) refers to one instance of the concept type. In our case, the concept type set S^L_{CG} includes the objects of the real world present in the photographs. They are application dependent and organized in a lattice L^L_{CG} that reflects generalization / specialization relationships. The relationship set S^R_{CG} includes absolute spatial, relative spatial and structural (like Comp, Label in the left part of Figure 7) relationships. Absolute spatial relationships link the image and the object concepts and indicate the position of the center of gravity of the regions sub-sampled to be an integer in [0,5].

The weighting scheme is inspired from [20], but we only consider media dependant weights. Compared to the *tf*idf* values as defined in [24] that models both the importance of a term in the document and with respect to the document collection, we limit ourselves to weights that compute visual term frequencies. However we input the certainty of the recognition of the concepts that is used in our representation. Hence we associate one concept with two values:

- The weight of the concept w_i, that represent the importance of the concept in the photograph. Many parameters may influence the weight of the objects. We compute the weight of an object as the probability that one pixel of the photograph may be in its region: w_i = surface(region$_i$)/surface(image).
- The certainty of recognition of the concept c_i. The results reported use certainty values that come directly from the VK labeling process.

A concept is then represented as a [*Type: referent* | w_i | c_i]. Figure 7 shows a part of a conceptual graph describing a real photograph. In this figure, the graph represents the content of the image IMG0232 that contains 2 regions: one corresponds to sky while the second one corresponds to mountains.

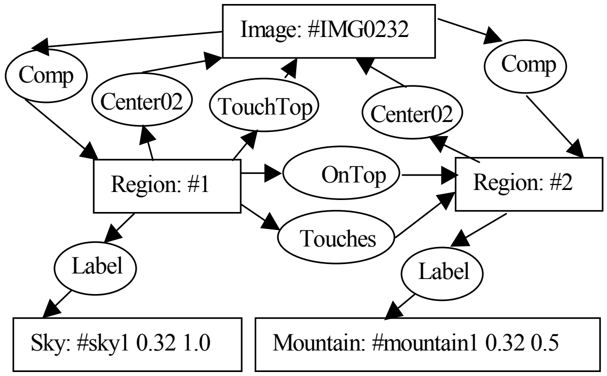

Figure 7. The conceptual graph of an image.

The query is also a conceptual graph generated from a natural language query. The query graph has the same components as the documents, except that the object labels are certain, have a weight of 1.0, and have only generic referents.

The matching process is two fold:

- We select the images that answer the query. This selection is based on the projection operators on a conceptual graph [28]. The projection operator intends to determine if the query graph is a sub-graph of an image graph, taking into account the lattices of concept types and of relations. Because a query graph may be projected into an image graph more than once, more than one projection may exist.
- We compute the relevance status value (RSV) for one query graph g_q and one document graph g_d. This matching value is computed according to the weights of the matching arches a_d (a component of a graph of the form [*Type$_{di}$: referent$_{di}$* | w_{di} | c_{di}] → (*Relation$_{dj}$*) → [*Type$_{dk}$: referent$_{dk}$* | w_{dk} | c_{dk}].) and the weights of the matching concepts c_d of g_d. The relevance status value for one query graph g_q and one document graph g_d is defined as:

$$RSV(g_d, g_q) = \max_{p_q \in \pi_{g_q}(g_d)}\Big(\sum_{c_d \text{ concept of } g_d} match_c(c_d, \pi c_d) + \sum_{\substack{a_d \text{ arch of } g_d \text{ that is corresponding to} \\ a_q \text{ of } p_q}} match_a(a_d, a_q)\Big)$$

where $\pi_{gq}(g_d)$ is the set of possible projections of the query graph into the image graph. The RSV formula may be compared to a dot product where the actual importance of arches and concepts of the image graph is one vector and the matching of the image parts and query parts forms another vector. The matching value *match*$_c$ between an image document concept c_d [$Type_{di}$: $referent_{di}$ | w_{di} | c_{di}] and a query concept c_q [$Type_{ql}$: $referent_{ql}$ | *1.0* | *1.0*] is computed as: $w_{di} \cdot c_{di}$. The matching value of arches is based on the importance of the considered arch [$Type_{di}$: $referent_{di}$ | w_{di} | c_{di}] → ($Relation_{dj}$) → [$Type_{dk}$: $referent_{dk}$ | w_{dk} | c_{dk}] of the document: $min(w_{di} \cdot c_{di,}\ w_{dk} \cdot c_{dk})$; this value is inspired from an fuzzy logic interpretation of conjunction. We remind that these values are computed only when we know that the query graph has at least one projection into the image document graph, so we ensure the meaning of the computed values.

2.4 Integrated Indexing

The indexing process, as presented on Figure 1, is a bottom-up task. The intermediate representation, *Fuzzy Object Map* (FOM), as depicted in Figure 5, is used to generate fuzzy labels for object regions based on a clustering algorithm in the fuzzy object space (see [15] for details). At the graph-based representation level, the most probable label is kept, and inference allows keeping specificities of relationships like symmetry and transitivity. As described previously the label set S^L_{VK} is the set of labels for the VK representation and S^L_{CG} is the set of concept types of the hierarchy of CG. The link between the indexes enforces $S^L_{VK} \subseteq S^L_{CG}$. Then, by keeping the two levels of representation, we are able to take into account recognition uncertainties as well as relationships among image elements.

2.5 Integrated Retrieval

The retrieval process is based on a combination of the VK and CG approaches. We combine the normalized relevance status values coming from the VK and CG sub-systems. Such combinations have been shown [2, 21] to be effective for text retrieval. The combination used is a weighted extension of the summation "CombSUM" [12], considering the normalized relevance status values of both VK (RSV_{norm_VK}) and CG (RSV_{norm_CG}):

$$RSV = \alpha\, RSV_{norm\_VK} + (1-\alpha)\, RSV_{norm\_CG}$$

We used $\alpha = 0.8$ in our experiments below which was determined via empirical tuning.

3. Experiments

Our experiments were conducted on a set of 2400 genuine family photographs collected over a period of 5 years. Figure 8 (on last page) displays typical photographs in this collection and Figure 9 (on last page) shows some of the photographs with inferior quality (left to right): fading black and white, flashy, blur, noisy, dark, and over-exposed photographs. These inferior quality photographs could affect any automatic indexing system but they were kept in our test collection to reflect the complexity of original and realistic family photographs. We focus on actual family photographs instead of the more general image collections like Corel images because our research aims to create useful automatic tools for mass consumers to organize and retrieve their home photographs. Moreover, we are convinced that dealing with real home photographs taken by average consumers is more challenging than working with professional stock photographs (like the Corel collection) used by many other researchers (e.g. [30]) because the quality and content of home photographs are more varied and heterogeneous. In fact, experiments reported in [30] confirmed that classification and retrieval results for the amateur home photographs were on the whole worse than those for the Corel images.

Table 1. The queries defined on the family collection.

Q1: Close-up of people	VOL
Q2: Small group of people at the center	CRL
Q3: Large group of people at the center	CRL
Q4: Any people at the center	CRL
Q5: Close-up of people, indoor	VOL
Q6: Small group of people, indoor	VOL
Q7: Large group of people, indoor	VOL
Q8: Any people, indoor	CRL
Q9: Any people	CRL
Q10: People near/besides foliage	CRL
Q11: People between foliage	CRL
Q12: People near building (or artifacts)	CRL
Q13: People in front of building (or artifacts)	CRL
Q14: People in front of mountain/rocks	CRL
Q15: People between water	CRL
Q16: People on one side of water	CRL
Q17: Flowers in a garden	VO:L
Q18: In a park or on a field	VOL
Q19: Close-up of building	VOL
Q20: Road/street scene in a city	VOL
Q21: Cityscape (view from far)	VOL
Q22: Mountain, view from far	VOL
Q23: At a (swimming) pool side	VOL
Q24: Object at the center, indoor	CRL

The 2400 photographs were indexed automatically as described in Section 2. From only 53 images (i.e. 2.2% of the whole test collection), we define and label 85 visual keywords. The detection of faces in the photographs was further enhanced with specialized face detector [23]. The overall number of labels for VK and CG are thus 85 and 110 respectively. CG also uses 48 relations, and applies symmetry and transitivity among relations when needed.

We defined 24 queries and their ground truths among the 2400 photographs (Table 1): the queries cover a wide range of potential queries, like close-up portraits (queries Q1, Q5), relative locations of objects (queries Q10, Q12 etc), absolute location of objects (queries Q2-4, Q24), and generic concepts ("large group of people", "indoor", "object" etc). We indicate for each of the queries listed in Table 1 if it is based mostly on the Visual Object Level (VOL) or the Concept-Relation Level (CRL) as described in the introduction.

Consider the 11 queries corresponding to the Visual Object Level:

- The query Q1 is only related to the occurrence of a face,
- The queries Q5, Q6 and Q7 look for close-ups of people and also on groups of people in the context of indoor photographs by looking into indoor objects.

- The other queries (Q17-Q23) consider the occurrences of objects like flowers, foliages, road, buildings, mountain and water.

Consider now the 13 queries corresponding to the Concept-Relation Level:

- The queries Q2 and Q24 and involve spatial relationships between elements,
- The queries Q4, Q8 and Q9 are based on the use of a generic concept People that correspond to any kind of people (face of a close-up, people in small and people in large groups),
- The queries Q10 to Q16 involve any people with specific relationships with other concepts (any kind of water for instance).

For each query listed in Table 1, we selected 3 relevant photographs as QBE input to the VK subsystem and constructed relevant textual query terms as QBS input to the CG subsystem. The query processing for QBE/VK, QBS/CG, and RSV integration are carried out according to subsections 2.2, 2.3, and 2.5 respectively.

We first discuss the reason why the VK results are more precise than those of the CG. The CG approach represents only the most probable label of an element recognized in a photograph. When the indexing is manually assisted we have very accurate descriptions. But here, the indexing is automatic and errors are inevitable, and this leads to the lack of precision of the CG approach in our results. On the contrary, the VK approach preserves the fuzziness of the labels during similarity matching and this helps to increase the quality of the results. The HSV Local method provides good results, but less accurate than those of the VK, and much less than the combined method.

We focus first on the general results related to the experiment. The Table 2 presents the average recall/precision results (over the 24 queries) obtained for the color-based indexing method, for VK and CG only respectively, and for the best combination of both. The first method, namely, “HSV Local”, can be seen as a special case of visual keywords method. The visual keywords chosen are eleven key colors (red, green, blue, black, grey, white, orange, yellow, brown, pink) in the HSV color space adopted by the original PicHunter system [5]. The indexing of “HSV Local” was carried based on the tessellation shown in Figure 6. Hence the “HSV Local” method is equivalent to locally weighted color histograms. The “HSV Local” method is only presented here to compare the results of usual non-symbolic approaches to our work. In Table 2, the values in brackets in the last column indicate the relative difference between the combination versus the individual HSV Local, VK and CG results, respectively. For the recall/precision results for VK and HSV Local, we removed the 3 query images from the relevant set to be fair.

The combined result shows that our integrated approach outperforms each of the individual subsystems by 8.1% to 32.2% (last row of Table 2). We also notice that significant improvements of the precision values happen at low recall values. This is very important for a practical system to be used by home users. For the query Q11 for instance, “people between foliage”, the average precision is 0.56 for the VK and 0.36 for the CG, but the combination provides an average precision of 0.62 (+10.7% over VK). This shows that, to a great extent, when the query is general, the query by example process is not appropriate and the use of a higher-level representation such as CG that includes hierarchies of relevant concepts is useful even if the CG representation may not be perfect.

Table 2. Recall/Precision table.

Recall	HSV Local	VK	CG	Combined
0	0.969	0.969	0.959	0.969 (0.0%, -0.6%, +1.0%)
0.1	0.503	0.659	0.453	0.705 (+40.1%, +7.0%, +55.6%)
0.2	0.367	0.457	0.345	0.515 (+40.2%, +12.5%, +49.2%)
0.3	0.304	0.365	0.304	0.392 (+28.9%, +7.5%, +28.8%)
0.4	0.278	0.290	0.248	0.323 (+17.1%, +11.4%, +29.9%)
0.5	0.254	0.261	0.212	0.279 (+9.7%, +6.8%, 31.4%)
0.6	0.234	0.237	0.198	0.251 (+7.6%, +6.2%, +26.8%)
0.7	0.214	0.215	0.187	0.229 (+7.3%, +6.3%, +19.0%)
0.8	0.197	0.194	0.175	0.205 (+3.9%, +5.5%, +17.2%)
0.9	0.179	0.176	0.164	0.182 (+1.7%, +3.5%, +10.9%)
1	0.141	0.141	0.138	0.044 (-6.9%, -6.9%, -6.8%)
Avg. Prec.	0.305	0.334	0.273	0.361 (+18.4%, +8.1%, +32.2%)

To be more precise about the results for low recall values, Table 3 presents the average precision values at the top 20, 30, 50, and 100 retrieved images for the HSV Local, VK, CG and the integrated approach. We chose these recall points comes from the fact that for image retrieval, the system is able to present 20 or 30 image thumbnails per page (or screen), and browsing a set of 50 or 100 images is not too painful for users (for instance, the well known Google web retrieval system (http://www.google.com) presents by default 20 query results for images, and 10 query results for text).

Table 3. Precision at 20, 30, 50 and 100 documents.

	HSV Local	VK	CG	Comb.
Avg. prec. at 20 images	0.493	0.592	0.408	0.621
Avg. prec. at 30 images	0.433	0.517	0.363	0.569
Avg. prec. at 50 images	0.378	0.437	0.330	0.488
Avg. prec. at 100 images	0.323	0.351	0.305	0.403

Once again, from Table 3, we observe that our combined approach outperforms all the other methods. In particular, on average, the combined approach retrieves between 0.6 and 4.3

more relevant images among the top 20 images. Applying the same argument to the 30, 50 and 100 retrieved image precision values implies that the combined approach finds between 1.6 and 6.2 more relevant images in the first 30 retrieved images, between 2.6 and 7.9 more relevant images in the first 50 retrieved images, and 5.2 and 9.8 more relevant images in the first 100 retrieved images. In short, our experimental results show that by combining the VK and CG approaches to take advantage of different levels of indexing representations, we achieve better retrieval performance over individual approach for heterogeneous queries on a large home photograph set.

Last but not least, we illustrate a sample retrieval result for query Q10 (People near/besides foliage) using our combined approach in Figure 10 (on last page). The retrieved photographs are displayed in the decreasing order of RSV in top-down, left-to-right manner. In this case, all the top 18 photographs returned by the system are relevant.

4. CONCLUSION

In this paper, we have presented a multi-level indexing and retrieval approach for photographic image retrieval. We differentiate the elements taken independently and their relationships in two indexing levels. The link between these levels, realized by the Visual Keywords [13-15] and the Conceptual-Graph approaches [17, 20], is proposed. We studied different query modes and the links among them as well as the links to the indexing levels during retrieval. Last but not least, we demonstrated the effectiveness of our proposed approach on 2400 home photographs with 24 queries.

However for each query in our experiments, a user has to issue a QBE with 3 query images and construct a QBS that expresses the relevant concepts and relations. It should be possible to invoke only QBE for the VK subsystem (or QBS for the CG subsystem) and automatically generate the QBS for the CG subsystem (or QBE for the VK subsystem). In the future, we will explore in greater depth the interaction between indexing and query processing levels as well as learning of semantic classes from examples and domain knowledge. A theoretical model for integrating different indexing representations and query processors will also be defined.

5. REFERENCES

[1] J. Bach, C. Fuller, A. Gupta, A. Hampapur, B. Horowitz, R. Humphrey, R. Jain, and C. Shu. The Virage image search engine: An open framework for image management. In Proc. SPIE Storage and Retrieval for Image and Video Databases, vol. 2670, pp. 76-87, USA, 1996.

[2] N. J. Belkin, C. Cool, W. B. Croft, J. P. Callan. Effect of multiple query representations on information retrieval system performance. In Proc. of SIGIR'1993, pp. 339-346, USA, 1993.

[3] A. Del Bimbo. Visual Information Retrieval. Morgan Kaufman Publisher. 1999.

[4] B. Bradshaw. Semantic-based image retrieval: a probabilistic approach. In Proc. of ACM Multimedia'2000, pp. 167-176, USA, 2000.

[5] In Cox, M. Miller, T. Minka, T. Papathomas, and P. N. Yianilos, The Bayesian Image Retrieval System, PicHunter: Theory, Implementation and Psychophysical Experiments , IEEE Transactions on Image Processing - special issue on digital libraries, (to appear).

[6] C. Faloutsos, R. Barber, M. Flickner, J. Hafner, W. Niblack, D. Petkovic and W. Equitz.. Efficient and effective querying by image content. Journal of Intelligent Information Systems, 3, 231-262, 1994.

[7] M. Flickner, H. S. Sawhney, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele and P. Yanker. Query by image and video content: the QBIC system. IEEE Computer, 28(9), 23-30, 1995.

[8] A. Gupta. Visual information retrieval technology: A Virage perspective. Virage Image Engine API Specification, Revision 4, Virage Inc., Feb. 11 1997.

[9] A. Gupta, T. Weymouth, and R. Jain. Semantic queries with pictures: The VIMSYS model. In Proc. of VLDB'91, pp. 69-79, Spain, 1991.

[10] C. Jörgensen. Attributes of Images in Describing Tasks. Information Processing and Management, 34(2/3): 161-174, 1998.

[11] J. Koenemann and N. J. Belkin. A case for interaction: A study of interactive information retrieval behaviour and effectiveness. In Proceeding of Computer Human Interaction Conference, pages 205-212, 1996.

[12] J. H. Lee. Analysis of Multiple Evidence Combination. Proc. Of the 20th ACM SIGIR, Philadelphia, pp. 267-276, USA, 1997.

[13] J.H. Lim. Explicit query formulation with visual keywords. In Proc. of ACM Multimedia'2000, Los Angeles, pp. 407-409, USA, 2000,.

[14] J.H. Lim. Building visual vocabulary for image indexation and query formulation. Pattern Analysis and Applications (Special Issue on Image Indexation), 4(2/3): 125-139, 2001.

[15] J.H. Lim. Fuzzy object patterns for visual indexing and segmentation. In Proc. of FUZZ-IEEE'2001, Melbourne, 2001.

[16] W. Y. Ma and B. S. Manjunath, NETRA: A toolbox for navigating large image databases. In Proc. IEEE ICIP'97, Santa Barbara, pp. 568-571, USA, 1997.

[17] M. Mechkour. An extended model for image representation and retrieval. In Proc. of the Intl. Conf. on Database and Expert System Applications (DEXA'95), pp. 395-404, UK, 1995.

[18] "Overview of the MPEG-7 Standard", ISO/IEC JTC1/SC29/WG11 N4031, Singapore, march 2001, http://mpeg.telecomitalialab.com/standards/mpeg-7/mpeg-7.htm

[19] P. Mulhem and L. Nigay. Interactive information retrieval systems: From user centred interface design to software design. In Proc. of ACM SIGIR'96, pp. 326-334, Switzerland, 1996.

[20] I. Ounis and M. Pasça. RELIEF: Combining expressiveness and rapidity into a single system. In Proc. of the ACM SIGIR'98, pp. 266-274, Australia, 1998.

[21] T.B. Rajashekar and W.B. Croft. Combining automatic and manual index representations in probabilistic retrieval. JASIS, 46(4): 272-283. 1995.

[22] S. Ravela and C. Luo, Appearance-based global similarity retrieval of images. In Advances in Information Retrieval: Recent Research from the Center for Intelligent Information Retrieval, W.B. Croft (Editor), pp.267-303, Kluwer Academic Publishing, 2000.

[23] H.A. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. IEEE Trans. on PAMI, 20(1), pp. 23-38, 1998.

[24] G. Salton G. and M.J. McGill. Introduction to Modern Information Retrieval, McGraw Hill, 1983.

[25] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain.Content-based retrieval image retrieval at the end of the early years. IEEE Trans. on PAMI, 22(12), pp.1349-1380, 2000.

[26] J. Smith and S.-F. Chang, Content-based image query system. In ACM Multimedia'96, pp. 2670, 1996.

[27] J. Smith and S.-F. Chang. Tools and techniques for color image retrieval. In IST & SPIE Proc. Storage and retrieval for Image and Video Databases IV, vol. 2670, pp. 87-98, USA, 1996

[28] J. Sowa. Conceptual Structures: Information Processing in Mind and Machine. Addison-Wesley Publisher, 1984.

[29] J. Sowa. Knowledge Representation: Logical, Philosophical and Computational Foundations. Brooks/Cole Publisher, 2000.

[30] C. Town and D. Sinclair. Content-based image retrieval using semantic visual categories. Technical Report 2000.14, AT&T Laboratories Cambridge, 2000.

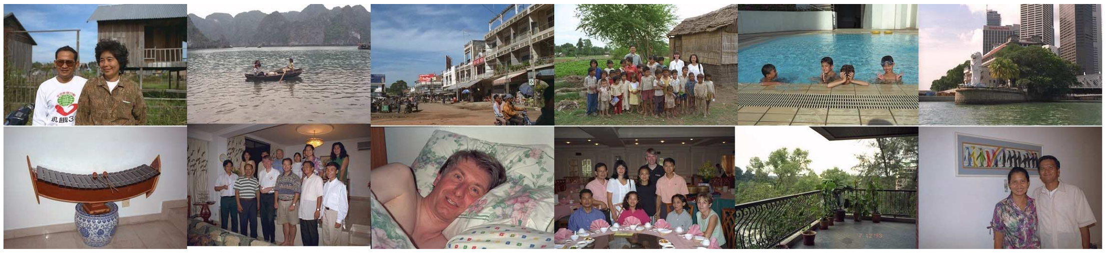

Figure 8. Typical photograph used in our experiments.

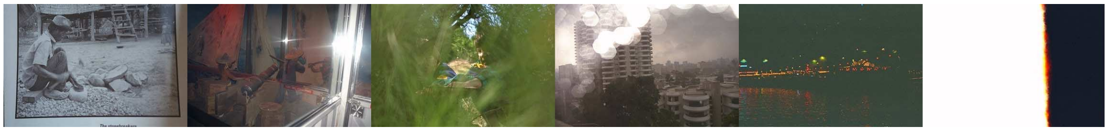

Figure 9. Inferior quality photographs used.

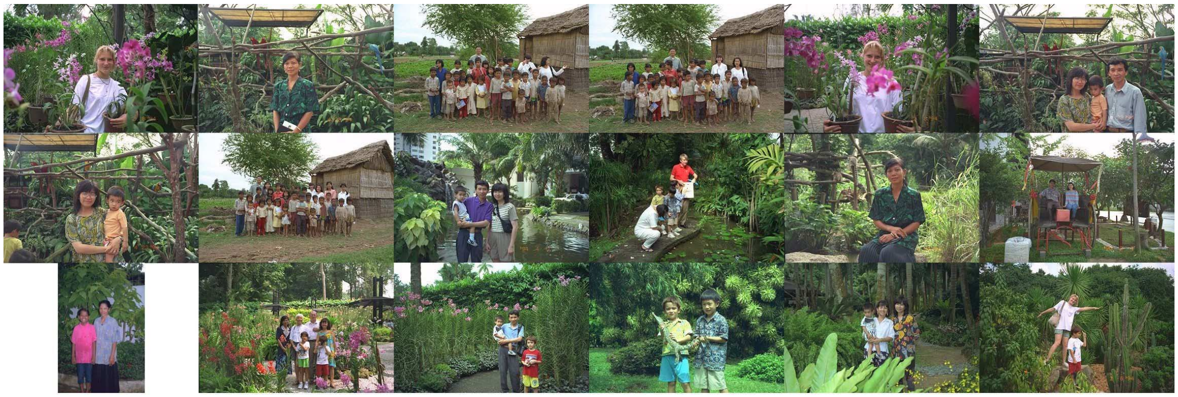

Figure 10. Top 18 images for Q10 (People near Foliage).