Advances in Digital Home Photo Albums

Philippe Mulhem[1], Joo Hwee Lim[2], Wee Kheng Leow[3], Mohan S. Kankanhalli[3]
1. IPAL-CNRS, Singapore
2. Laboratories for Information Technology, Singapore
3. School of Computing, National University of Singapore

1. Introduction

The last few decades has witnessed a dizzying rate of technological innovation in the areas of computing and communication. While the effects of cheaper and faster computation power manifest explicitly in many places, slightly less obvious are the technological advances in sensor and signal processing technologies. Their impact is being increasingly felt in the form of digitization of all forms of communication and media. These advances have directly led to increased inexpensive communication bandwidth, which in turn has spurred the rapid acceleration of the Internet globally. In terms of consumer electronics devices, we are currently witnessing the mass-scale switchover to digital cameras from the traditional analog cameras. In fact, the year 2002 is expected to be a significant milestone since it will perhaps become the first year when the sales of digital cameras will outstrip those of their analog predecessors.

In fact, the history of the invention of the camera is quite interesting. Joseph Nicéphore Niépce and Louis-Jacques-Mandé Daguerre's invention of photography in 1825 was followed by the landmark contributions of George Eastman's dry photographic film with the associated camera in 1888 and Edwin Land's instant Polaroid photography in 1948. Essentially, the camera has democratized the preservation of images and thus it has had a tremendous impact on society. In the ancient times, only the kings and nobles could afford to engage artists to paint portraits, depict monuments and glorify conquests. With the invention of the camera, even ordinary people's lives started getting captured visually through this powerful medium. It basically eliminated the need of a skilled intermediary and simultaneously collapsed the time-period between the intent and production of a visual memory. Thus, the camera has been one of the first technological devices to be utilized on a large scale by people that demand neither the mastery of the technology nor the refinement of artistic skills. Hence it rapidly became a mass-market consumer device. We are in fact at an interesting technological cusp today. On one hand, the falling price of digital devices is rapidly pushing up the sales of digital cameras. On the other hand, increased global affluence coupled with the growing mobility of people is leading to an ever-greater use of the camera. Consequently a huge amount of digital images is being generated everyday.

We are quite familiar with the traditional paradigm of handling photographs. Analog cameras are used with a film roll that can capture several photos, which are processed at one shot. The output of the rolls used for significant events such as birthdays, weddings, graduation ceremonies and travel are stored in a *home photo album*. This paradigm of

capturing memories strongly resembles the book paradigm and quite easy to use for most people. Though it has some disadvantages in terms of limited use capabilities in terms of making copies or searching by only global album labels, it is a familiar and comfortable approach for most people.

Given that we are undergoing a paradigm shift in terms of the camera technology, we can respond to this change in two ways. One way of responding to the challenge of managing large numbers of digital photographs is by faithfully mapping the analog photo album paradigm onto the digital arena, replicating both the look and functionality of the traditional approach. An alternative response would be totally rethink and completely reengineer the way home users create, store, and manage *digital home photo albums*.

Thus, with more and more digital photos being accumulated, home users definitely need effective and efficient tools to organize and access their images. To address this genuine need, a three-year international collaborative project between CNRS, France, School of Computing, National University of Singapore, and Laboratories for Information Technology, Singapore was formed in 2000 to develop the next generation Digital Image/Video Album (DIVA) for home users. The goal of this chapter is to explain the work carried out in this DIVA international project. In particular, we will describe the needs of home users and possible solutions in the domain of management of digital images.

At the very outset, the DIVA project has adopted the second approach of having a fresh look at the digital home photo albums in order to come up with an appropriate solution without being encumbered by habits and legacies of past. Having settled on the approach, we came up with some fundamental assumptions related to digital home photo albums:

- A digital home photo album should not engender any *digital divide* between the ordinary home user and a technologically savvy home user. In particular, the digital home photo album should be intuitively easy to use for most users while offering the subtle flexibilities to the sophisticated user.
- Users take photographs (digital or otherwise) in order to *preserve* and *share* memories across space and time. Thus, sharing of home photos is at least as important as archiving the photographs.
- Given that sharing is important, users should be provided with extremely flexible means of *searching* and locating the appropriate photographs that are to be shared. This implies that the semantic content of the photographs needs to be represented along with the image.
- The user is likely to share his/her photographs with various people of different degrees of acquaintance and varying tastes. Hence each presentation of shared photographs should be different - tailored to that particular individual or target group. Hence, the ability to build *custom presentations* out of the same set of photographs is absolutely necessary.
- The digital home photo album of a user should adapt itself to the quirks and predilections of the user (instead of vice versa). For example, common content of the images such as the user's face, name and family members should be learned

by the album over a period of time instead of being a memory-less stateless system. Thus, implicit and explicit *personalization* should be built into the system.

Given the above set of fundamental assumptions, we need to develop several technical pieces before we can solve the technological jigsaw puzzle. In particular, these technical considerations stemming from the fundamental assumptions need to be addressed:

1. Image capture and transfer: how should the image (and perhaps the camera parameters such as focal length, flash use etc.) captured on the digital camera be transferred to the digital home photo album which could reside on a computing device like a personal computer or possibly an embedded system consumer appliance like a portable image juke-box or perhaps be integrated with the camera itself.
2. Image coding techniques: for efficient storage of the digital image.
3. Image annotation: how to index the images with information related to the semantic content of the image. In other words, capturing the metadata about the digital photo.
4. Representation of the metadata: which can facilitate flexible searching.
5. Query languages: in order to pose the search constraints which enables efficient searching.
6. User interfaces: which obliterates any digital divide by making the use very natural.
7. Presentation tools: in order to customize the presentations for target audiences.
8. Web interface: in order to use the worldwide web as a global channel for sharing of images.
9. Personalization: the system should have means of personalizing explicitly the layout, the metadata and the sharing mechanisms. Moreover, it should strive for implicit personalization based on observed user behavior.

Thus, how to manage digital images is a challenging problem. We will now detail some of the efforts towards resolving a few of the above issues in the DIVA project. While we are taking an integrated approach of solving this problem in our project, our focus in this chapter is biased by the overall theme of this book which basically deals with the content-related aspects. Hence we will discuss in detail the content-based retrieval aspects while only briefly alluding to the other aspects.

2. Users needs

Before going in detail into the existing work in content-based photographic images retrieval, we focus on the needs of users for the management of their home photographs. Even if the role of home photographs, i.e. a "bank of memory" [Chalfen 1998], is somewhat clear, we had to get users point of view about how home photographs management systems may help them.

The results presented here were obtained from a study carried out in 1999 on a panel of 37 persons. We intended to find out the current interest of users when organizing their images, and the consumer's expectations about the future home photo management software.

Figure 1 shows how people currently retrieve and organize their photographs. In this figure, we present for each of the photographic points of view the relative ranked importance given by users. The points of views are: the *presentation* related to the colors/textures/shapes present in the images, the *content* related to the objects present in the images, their actions and their relations, the *context* related to the description of the date and the event at the origin of the photograph, and *connotation* related to the mood associated to the photograph. According to the figure 1, we see that the *context* is the most important point of view to organize and retrieve images, and the *content* also plays a role in such photographs organization.

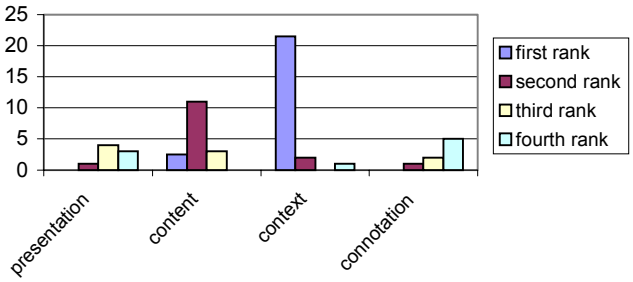

Figure 1. Existing means to organize still images.

The second step is more related to study the potential uses expected by users for the future home photograph systems. The question was for users to rank the expectations from first to fifth. The expectations proposed were: retrieval, organizing, duplication of photographs, and gain of space of digital systems versus the paper. As we see in figure 2, the users expectations are first related to retrieval and second to organization of photographs. These results are consistent with the findings of [Rodden 1999]. We conclude then that future home photo systems are expected to be able to handle queries and to help the user to organize her/his visual data. Organizing such visual data includes the creation of presentations like slide shows. FlipAlbum Suite [FlipAlbum 2002] proposes to mimic the paper-based photo albums for presentation, whereas 3D-Album [3D-Album 2002] considers various 3D styles for presenting photographs.

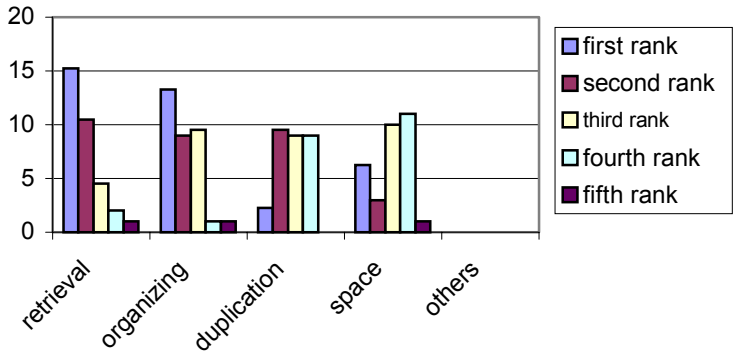

Figure 2. Expectations for home photo management systems.

Because the retrieval and organization of the images are the most important aspects related to the management of home photographs, we consider in the next section the related works in the domain of still images indexing and retrieval before explaining or approaches based on symbolic descriptions of the image content (considered very important by users).

3. Related work on content based image retrieval

Indexing and retrieval of images is a research field that has been studied for more than a decade. We describe the existing literature based on a simple classification that considers ascending approaches as descriptive approaches that go from the signal (i.e., signal-based), descending approaches more related to the conceptual representation (i.e., symbolic-based), and mixed approaches that use both symbolic and image processing techniques, and interpretation-based approaches.

At present [Smeulder et al. 2000], the overwhelming majority of existing approaches related to image retrieval are descriptive. For instance, consider QBIC [Faloustos et al. 1994, Holt et al. 1997], VisualSeek [Smith et al. 1996a, 1996b], Virage [Gupta 1995, Bach et al. 1996], and Netra [Ma et al. 1991]. All of these approaches use color, texture and shape features to represent the content of images, but they never fill the gap between the description of objects and the actual symbols that correspond to the objects. Such approaches have the great advantage of avoiding promising impossible tasks: the user has to fill the gap between the interpretation and the description of the desired images when she/he asks a query. If the user does not perform this task properly then the results of such systems cannot be guaranteed. On the other hand, the significant advantages of such systems are first related to the fact that they can be applied on many kinds of photographs because they use only little a priori features of image collections (except colors), and second that they are almost fully automatic.

Mixed Approaches integrate human interaction and image processing techniques. For instance, the MiAlbum [Wenyin et al. 2000, Lu et al. 2000] uses relevance feedback

sessions from users to "infer" the link between the features and keywords. MiAlbum counts on the number of relevance feedbacks (from a web site) to obtain accurate indexes. The Photobook [Minka and Picard 1997] approach also uses interaction to propagate symbolic keywords along the image collections. For the still images, our approach is more related to the mixed approaches. We consider that face recognition belongs to this class, because the models of faces have to be learned by the system according to some input (face collection). Other works like [Bradshaw 2000], [Town and Sinclair 2000] or [Luo and Etz 2002] are also based on more or less complex approaches to be able to apply symbolic labels to images or image regions.

Descending approaches consider only the interpretation of images. For instance, the EMIR2 model [Mechkour 1995] used also in the FERMI-CG work [Ounis & Pasça 1998] focus on the representation of the interpretation of the image content in term of objects and relations of objects using a knowledge representation formalism, the conceptual graphs [Sowa 1984]. The Symbolic Description of MPEG-7 [MPEG-7 2001] also presents a modeling framework for objects and their inter-relations that can be applied on images. The advantage of such approaches is that they directly tackle the symbolic content of images, and are able to represent information that cannot, according to the state of the art in image analysis and computer vision, be extracted automatically. They also consider information that deals with purely symbolic characteristics (like generic/specific relationships between objects, or part-of relationships). It is well known however that such manual indexing data is not very consistent across human beings, and so great care has to be taken during such an indexing process. Another major drawback of such descending approaches is that the indexing of images is then a tedious process. However, speech analysis may ease the indexing drudgery and aid the retrieval process, like with Show&Tell [Srihari and Zhang 2000], Shoebox [Mills et al. 2000], and SmartAlbum [Chen et al. 2002] systems. Sophisticated interaction techniques, like *drag-and-drop*, used in the PhotoFinder project [Schneiderman and Kang 2000] or propagation techniques like in FotoFile [Kuchinski et al. 1999] can also help significantly. Other approaches intend to use image metadata coming for the camera itself, like [Gargi et al. 2002] using EXIF [EXIF 2.0 1998] information stored with photographs taken by consumer digital cameras. In this case, the information extracted is numerical, and mixed approaches are devoted to detect and recognize faces. Considering one specific kind of meta-data, namely the date of creation of the photograph, the Calendar Browser [Graham et al. 2002] clusters images and generates a hierarchy to ease the navigation in the consumer image database.

From another perspective, efforts have also been made to develop ways for a consumer to easily browse through his collection of photographs. The PhotoFinder system [Kang and Schneiderman 2000] presents a browser with a zooming function to provide continuity during the browsing. Fotofile [Kuchinski et al. 1999] proposes hyperbolic trees to rapidly access images.

In the context of home photographs, as we saw in section 2, the symbolic content of images is much more important than low-level features, that is why our work focusses on such symbolic representations of images. Because a home photo management system has

to be easy to use by consumers, it as to be as automatic as possible, that is why we consider mixed approaches that try to extract the symbolic concepts from the image itself.

4. Digital Home Photo Album Indexing and Retrieval

We propose a new methodology to semantically index and retrieve photos. We advocate a learning-based approach that allows creation of semantically meaningful visual vocabularies (e.g. faces, crowd, buildings, sky, foliage, water etc) from examples. These semantic visual representations are further spatially aggregated and organized into conceptual graphs for describing the concepts and relations in the image contents. The home users can therefore query the system by visual example(s), by spatial arrangement of visual icons (i.e. a visual query language), and by concepts and relations in text, beyond the conventional indexing and retrieval based on low-level features such as colors, texture, and shapes.

4.1. Symbolic Labeling

As we explained in section 2, consumers think of images using symbols and not low-level features, even if this "high-level" representation is obviously related to the "low-level" pixels seen in the photographs. That is why a large amount of the work in the DIVA project has gone towards automatically associating symbolic descriptions to the photographs. We now explain the different steps that are involved in coming up with accurate descriptions of images. The two approaches used in the context of DIVA generate labels from the low-level image content through the use of probabilistic frameworks and by learning of visual keywords.

Symbolic labeling can be framed as a problem of classifying an image region (or image patch, fixed-sized block) R to one of several semantic classes C_i, $i = 1,\ldots, M$. In practice, it is not possible to achieve perfect classification due to the presence of noise and ambiguity in the region features. A better approach is to represent the uncertainty of classification in the semantic labels and to defer the final decision to a latter stage at which the image structure can be taken into account using, for instance, the fuzzy conceptual graph matching method [Mulhem et al. 2001].

Let us denote by $Q_i(R)$ the confidence of classifying region R into semantic class C_i. Then, the symbolic labeling problem can be defined as follows:

> Given a region R and M semantic classes C_i, $i = 1, \ldots, M$, compute the confidence $Q_i(R)$ that R belongs to C_i for each i.

A region R contains a set of features F_t of type $t = 1, \ldots, m$, each having a value v_t. Each feature F_t can contain more than one component, like color histograms, Gabor texture features, wavelet features, etc. The symbols F_t and v_t denote the *whole* feature and feature value instead of the individual component.

The standard method of computing $Q_i(R) = Q_i(v_1, \ldots, v_m)$ is the vector space approach: Regard each set of feature values as a vector $\mathbf{v} = [v_1, \ldots, v_m]^T$ in a linear vector space, and estimate Q_i using function approximation. It is well known that the various feature types are not independent of each other and the scales of the feature types are different. So, forming a linear vector space by assigning each feature type (or, more commonly, each feature component of each feature type) to an orthogonal dimension of the vector space is not expected to produce reliable results in general.

Instead, we adopt a probabilistic method of computing $Q_i(R)$ by estimating the conditional probability $P(C_i \mid v_t)$. The approach encompasses the following characteristics:

- It can make use of the dissimilarity measures that are appropriate for the various types of features [Leow and Li 2001, Puzicha et al. 1999] instead of the Euclidean distance.
- It does not require the use of weights to combine the various feature types.
- It adopts a learning approach that can adapt incrementally to the inclusion of new training samples, feature types, and semantic classes.

The probabilistic labeling method consists of two stages: (1) semantic class learning and (2) region labeling.

Semantic Class Learning

The goal of the semantic class learning stage is to determine the conditional probabilities associated with each semantic class. This is accomplished by first partitioning the space of each feature type into discrete regions of approximately uniform probability density. An adaptive clustering algorithm is applied to cluster a set of training sample regions R_j, each assigned a pre-defined semantic class C_i, in the space of each feature type:

Adaptive Clustering
Repeat
  For each feature value vt of each region,
    Find the nearest cluster Ωtk to vt.
    If no cluster is found or distance d(Ωtk, vt) ≥ s,
      create a new cluster with feature vt;
    Else, if d(Ωtk, vt) ≤ r,
      add feature value vt to cluster Ωtk.
  For each cluster Ωti,
    If cluster Ωti has at least Nm feature values,
      update centroid of cluster Ωti;
    Else, remove cluster Ωti.

The centroid of cluster Ω_{ti} is a generalized mean of the feature values in the cluster and the function $d(\Omega_{tk}, v_t)$ is a dissimilarity measure appropriate for the feature type t [Leow

and Li 2001, Puzicha et al. 1999]. In the current implementation, the following feature types are used:

- Adaptive color histograms:
  An adaptive color histogram is obtained by performing adaptive clustering of the colors in an image [Leow and Li 2001]. Since different adaptive histograms can have different number of bins and different bin centroids, Euclidean distance and Euclidean mean cannot be applied on adaptive histograms. Instead, the weighted correlation dissimilarity measure [Leow and Li 2001] is used to compute the dissimilarity between two adaptive histograms, which has been shown to give better performance than the Earth Mover's Distance [Leow and Li 2001]. For computing cluster centroids in region clustering, a special histogram merging operation is used, which has been shown to produce the correct mean of histograms [Leow 2002].
- MRSAR texture features:
  The usual dissimilarity measure for MRSAR is the Mahalanobis distance, and the Euclidean mean is used to compute cluster centroids.
- Gabor texture features:
  Weighted-mean-variance (WMV) as defined in [Ma and Manjunath 1996] is a good dissimilarity measure for Gabor texture features. For computing the cluster centroids, Euclidean mean is used.
- Edge histograms:
  Normalized edge direction and magnitude histograms as defined for MPEG7 [MPEG-7 2001] are extracted from the images. For these features, Euclidean distance and Euclidean mean are used for region clustering.

After the clustering has converged (or a fixed number of iterations has been performed), some training regions may remain unclustered. In this case, the cluster nearest to an unclustered training region can be expanded to include the region. This will result in clusters having different radii r_{tk}.

The clustering process produces a set of clusters Ω_{tk}, for each feature type t. After clustering, the conditional probability $P(C_i \mid \Omega_{tk})$ for semantic class C_i given cluster Ω_{tk} is estimated. Assuming that the distribution within each cluster is uniform, then $P(C_i \mid \Omega_{tk})$ can be estimated from the number of regions in the cluster:

$$P(C_i \mid \Omega_{tk}) = \frac{P(C_i, \Omega_{tk})}{P(\Omega_{tk})} = \frac{| C_i \cap \Omega_{tk} |}{| \Omega_{tk} |}$$

where $| \Omega_{tk} |$ denotes the number of regions in cluster Ω_{tk} and $| C_i \cap \Omega_{tk} |$ the number of regions in Ω_{tk} that belong to semantic class C_i.

To combine multiple feature types, we can determine the cluster combinations $\Psi(\tau, \kappa, n) = \{\Omega_{\tau(1), \kappa(1)}, \ldots, \Omega_{\tau(n), \kappa(n)}\}$, $\tau(i) \neq \tau(j)$ for $i \neq j$, that have high probabilities of associating to some semantic classes C_i:

$$P(C_i \mid \Psi(\tau,\kappa,n) = P(C_i \mid \Omega_{\tau(1),\kappa(1)},\ldots\Omega_{\tau(n),\kappa(n)}) = \frac{\mid C_i \cap \bigcap_l \Omega_{\tau(l),\kappa(l)} \mid}{\mid \bigcap_l \Omega_{\tau(l),\kappa(l)} \mid}.$$

The functions $\tau(l)$, $l = 1, \ldots, n$, denote a combination of feature types and $\kappa(l)$ a combination of cluster indices.

Region Labeling

After the learning stage, a region R can be labeled by determining the associated semantic classes. Given the region R which contains a set of feature values v_t, the nearest clusters that contain the feature values v_t, for each feature type t, are determined. Next, these clusters found are matched with the stored cluster combinations, obtained during the learning stage, that are associated with some semantic classes. The confidence measure $Q_i(R)$ can now be computed from the conditional probabilities of the matching cluster combinations $\Psi(\tau, \kappa, n)$:

$$Q_i(R) = \max_{\tau,\kappa,n} P(C_i \mid \Psi(\tau,\kappa,n)$$

Note that $Q_i(R)$ as defined in the above equation is no longer a conditional probability:

- The C_i's in the equation may be conditioned on different sets of feature types.
- While $\sum_i P(C_i \mid \Psi(\tau,\kappa,n)) = 1$ for each cluster combination $\Psi(\tau, \kappa, n)$, the sum $\sum_i Q_i(R) \neq 1$ in general.

Nevertheless, $Q_i(R)$ is derived based on conditional probabilities and is, thus, a good measure of the confidence that region R belongs to class C_i.

4.2. Symbolic Indexing

The symbolic labeling is the first step to obtain symbolic descriptions of images. The next step is to represent the content of images in a powerful and efficient representation. This is achieved using two different *points of views* of images: the use of visual keywords that integrates the ambiguous interpretation of regions, and the use of extended conceptual graph formalism that is able to support the hierarchies of concepts and complex relationships.

4.2.1. Visual Keywords

The Visual Keyword approach [Lim 2001] is a new attempt to achieve content-based image indexing and retrieval beyond the feature-based (e.g. [Faloustos et al. 1994]) and region-based (e.g. [Smith et al. 1996a]) 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 4 key components (square boxes) as shown in Figure 3. 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. Thirdly, the fuzzy detection results are registered as a *Fuzzy Object Map* (FOM) and further aggregated spatially (spatial summary) into a *Spatial Aggregation Map* (SAM). Last but not least, with visual thesaurus, the SAM can be further abstracted (keyword abstraction) to a simpler representation, *Concept Aggregation Map* (CAM).

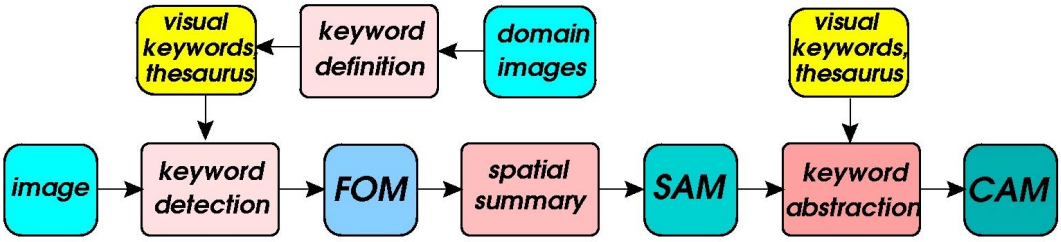

Figure 3. Workflow of the indexing process based on visual keywords

Visual keywords are visual prototypes specified or learned from domain-relevant regions of sample images. A set of labels $S^L_{VK}=\{C_i\}$ is assigned to these visual prototypes as a vocabulary. The labels C_i correspond to the semantic classes described in part 4.1. Figure 4 shows some examples of visual keywords used in our experiment.

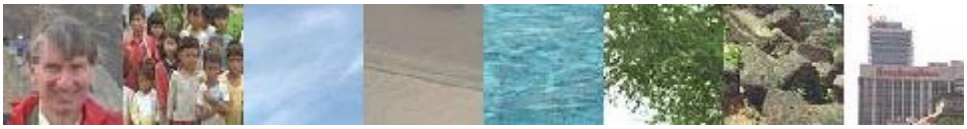

Figure 4. One visual keyword from each visual class: face, crowd, sky, ground, water, foliage, mountain/rock, and building.

An image to be indexed is scanned with square 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 [Lim 2001] for details). The Fuzzy Object Map actually represents the probabilities of the different labels for the considered window.

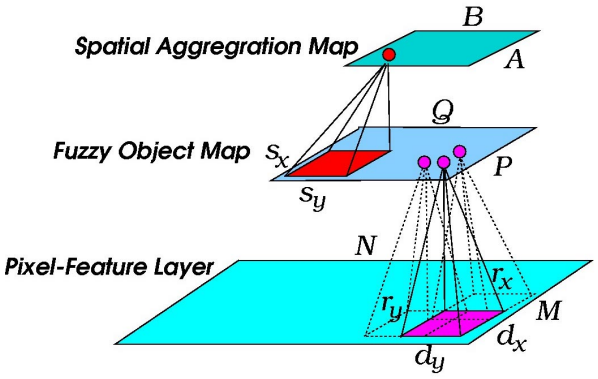

Figure 5. Automatic indexing with fuzzy recognition values for labeling

Visual keywords can be regarded as co-ordinates that span a new pseudo-object feature space. The scale on each dimension is the probability of that visual keyword being detected at a specific spatial locality in the visual content (Figure 6).

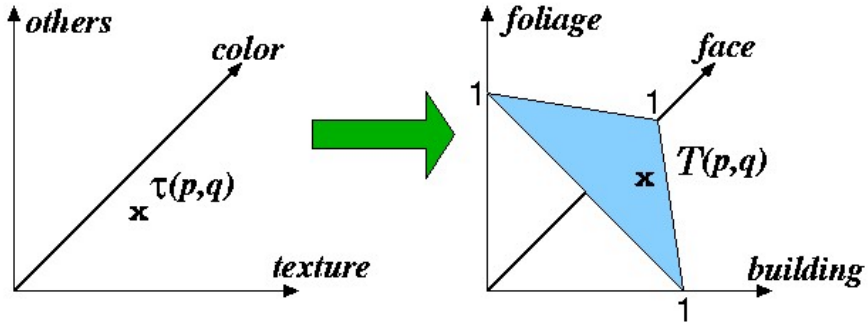

Figure 6. Transforming feature space into visual keyword space for a window of coordinates (p,q).

4.2.2. Extended Conceptual Graphs

To handle the needs of consumers in the context of home photo retrieval, we also propose the use of the Conceptual Graph (CG) formalism [Sowa 1984]. A conceptual graph is a directed bipartite graph composed of two kids of nodes: concepts and relations. In the initial definition of [Sowa 1984], a concept is itself composed by definition of a concept type and a referent. Concept types are organized in a hierarchy, representing for instance that the concept type Man is a specific of the concept type Human. In a concept, a referent can be generic, noted *, or individual, like *#john*. A generic referent only denotes any referent. The conformance relationship, namely Conf, exists between a concept type and individual referents that correspond to the concept type. For instance, Conf(Man,

#john) may exists. The figure 7 presents the alphanumerical and graphical notations for a concept of type Man and of referent *#john*.

[Man: #john]

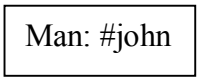

Figure 7. Aphanumerical (left) and graphical (right) notations of a concept.

Relations are also organized in a lattice reflecting genericity/specificity (reflecting for instance the fact that the relation "*drink*" is a specific of "*ingesting*", but they do not contain any referent. We consider only binary relations in the following, even if the CG formalism allows n-ary relations. The figure 8 shows the alphanumerical and graphical notations for a relation *ingest*.

(ingest)

Figure 8. Aphanumerical (left) and graphical (right) notations of a Relation

The figure 9 shows a graph representing "a man named John is walking on a road and is ingesting an apple" under it's alphabetical and graphical notations.

[Man: #john]->(walk)->[Road: *]
->(ingest)->[Apple: *]

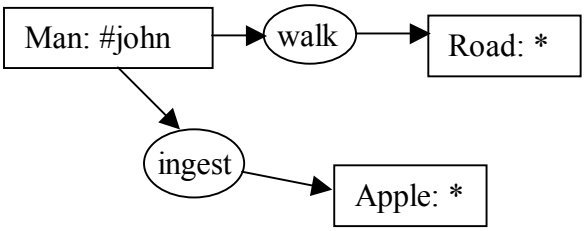

Figure 9. Aphanumerical (left) and graphical (right) notations of a Graph.

To ease the explanation following, we call a triplet (concept, relation,concept) of a graph an *arch*. For instance, the graph of figure 9 contains two arches.

To be able to fit onto an image retrieval model, we use (as in standard text information retrieval systems [Salton and Buckley 1988]) weights that represent the importance of the objects in the graph description. In our case, a weight is assumed to be directly dependant of the relative surface of the objects in the images. Another important element, specific to the fact that the recognition process (cf. section 4.1.) is not perfect, is to take into account the certainty of recognition in the concept representation.

This leads to the definition of an extension of the usual conceptual graphs, by taking into account the additional information related to conceptual graphs, namely the weights and certainty values. Figure 10 presents the alphanumerical and graphical description of a graph describing that a building *#b1*, having a weight of 0.4 and a certainty of recognition of 0.75, is at the left of a tree *#t1*, having a weight 0.2 and a certainty of recognition 0.9.

[Building: #b1 0.4 0.75]->(left_of)->[Tree: #t1 0.2 0.9]

Figure 10. Alphanumeric (top) and graphical (bottom) extended conceptual graph.

Having defined the conceptual graph formalism used to represent the content of images, we focus now on the model of images, i.e., the elements represented in this representation. The concepts types that are defined for the image model are composed of the objects that can occur in images, i.e. the set S^{L}_{VK}, organized in a hierarchy. Examples of concept types are "building", "foliage", etc. A concept type "Image" denotes one image, and the concept type "Region" denotes a region in the image. Based on these concept types, the overall structure of the image model expresses the fact that an image is composed, with the relation "Comp", of regions uniquely identified by a referent, each region being associated by a relation "Label" to an objet type uniquely identified by a referent. The other relations between the concept Image of an image and its Regions describe the spatial position of center of grapvity of the region (with the relations "center00", …, "center44"), and if the regions touch the border of the photograph (touch_top, touch_bottom, touch_right, touch_left). Between region concepts, relations expresses if the region touches (relation "touch") and the relative position of regions ("left_of", "on_top", "under", "right_of"). The figure 11 presents on its left part a photograph and on its right part an excerpt of the extended graph that describes its content. In this figure, only the weights and certainty values of the object are presented.

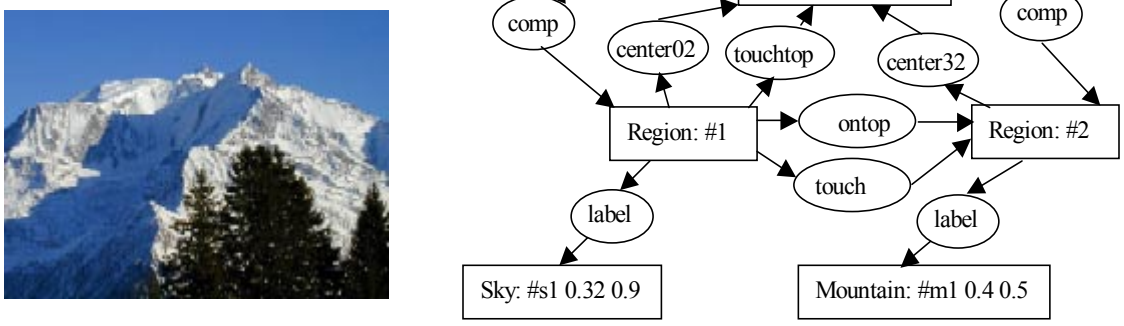

Figure 11. An example of the graph-based content representation for one image.

That is such representation of the photograph contents that are used during the retrieval based on graphs.

4.3. Retrieval

The retrieval of images is based on the two representations presented before: the visual keywords and the extended conceptual graphs.

4.3.1. Visual Keywords

As SAM (or CAM) summarises 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,

$$\lambda(x,y) = \frac{\sum_{(a,b)} \omega(a,b)\lambda(a,b)}{\sum_{(a,b)} \omega(a,b)} \qquad \text{and} \qquad \lambda(a,b) = 1 - \frac{1}{2}\left|SAM_x(a,b) - SAM_y(a,b)\right|$$

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 12.

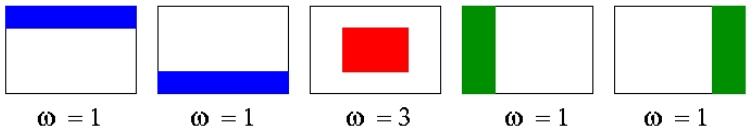

Figure 12. Tessellation and weights for similarity matching in QBE.

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.

4.3.2. Extended Conceptual Graphs

The retrieval process using conceptual graphs is based on the fact that a query is also expressed under the form of a CG. Without going into details, a simple grammar composed of a list of objects names and relations is easily translated into a graph, for instance the string: "people left of foliage" is translated into the graph of figure 13. In this figure, the concepts are all assumed to have a weight of 1 and a certainty of recognition of 1.

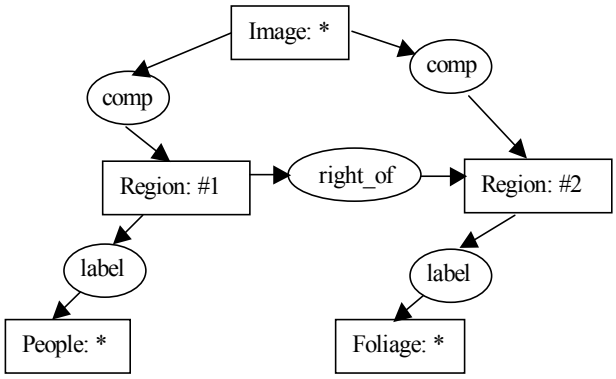

Figure 13. A Query graph example.

The matching process is two fold:

- First, we select the images that answer the query. This selection is based on the projection operators on a conceptual graph [Sowa 1984]. 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. If we consider the query example of figure 8, an image containing a photograph of John (a Man, and then a people) at the right a pine tree (a specific of foliage) will be selected because the query graph is included, according to the concept type and relation hierarchies, in the image graph. Because a query graph may be projected into an image graph more than once, more than one projection may exist.
- Second, we compute the relevance status value (RSV) for one query graph g_q and one document graph g_d. This relevance value computation is related to the fact that ranking query answers is a must for any information retrieval system, thus for image retrieval ones. 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:

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

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*] takes into account the fact that the weight and the certainty of recognition play both a role during retrieval, and is computed as: $\sqrt[\tau]{w_{di}} . \sqrt[1/\tau]{c_{di}}$. The parameter τ is defined experimentally. 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: $\min(\sqrt[\tau]{w_{di}} \cdot \sqrt[1/\tau]{c_{di}}, \sqrt[\tau]{w_{dk}} \cdot \sqrt[1/\tau]{c_{dk}})$; this value is inspired from a 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.

5. Still Photographs Management

As we emphasized previously, home photographs management systems are not only limited to the retrieval features. We present here two of the prototypes developed in the DIVA project: one, namely DIESKAU, concentrates more on the indexing/retrieval using extended conceptual graphs, while the second, namely PhotoCorner, concentrate retrieval, albums management and dynamic presentation tools.

The first one, DIESKAU (Digital Image rEtrieval System based on Knowledge representAtion and image featUres), allows precise descriptions of the targeted images using textual descriptions of images. The core of the system is based on an efficient implementation of the conceptual graphs formalism. The figure 14 presents a text query where a user looks for images containing people close to foliage and containing preferably a specific kind of building, namely a hut, touching one side of the photograph. The results presented in the right-hand side of the figure shows the results obtained for this query. The lower left part of the interface allows the user to define the different use of the parameters of the system during the retrieval process, namely putting more or less emphasize on the relations, concepts, certainty of recognition of importance. This interface allows also the use of query by example to retrieve images.

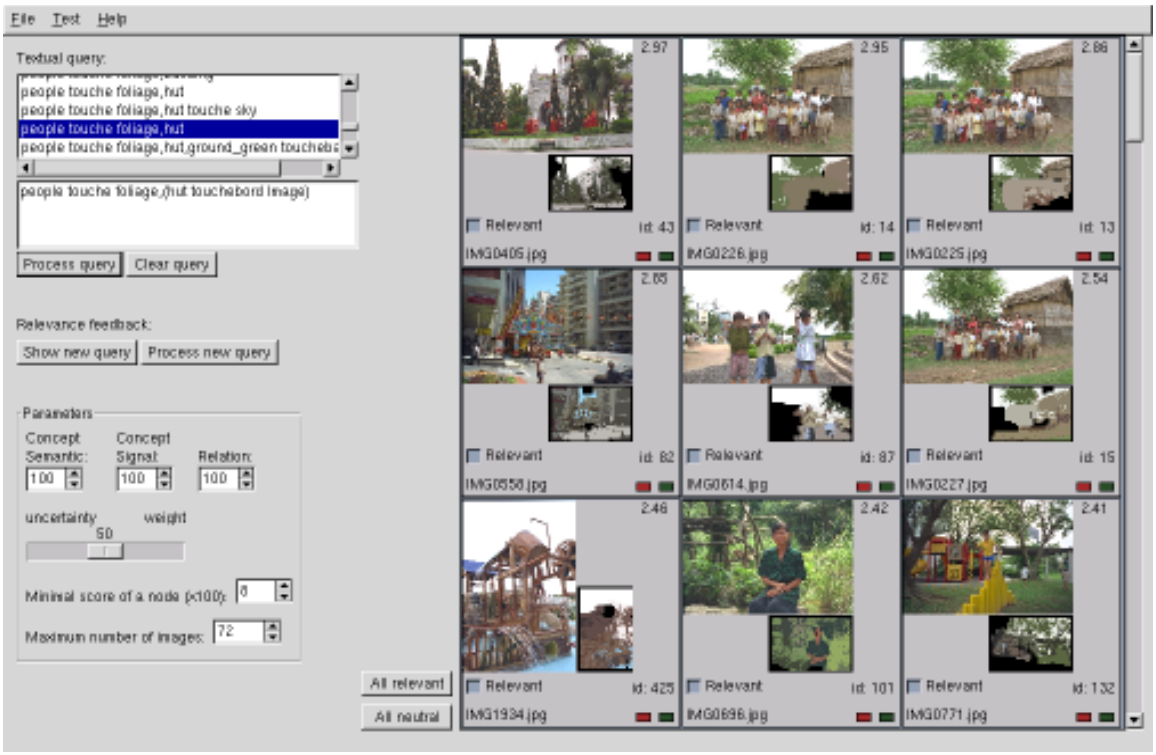

Figure 14. The DIESKAU Interface.

If the DIESKAU interface is useful for very precise query expressions or fine-tuning of the retrieval process, consumers prefer more simple ways to retrieve images. That is why we describe a web-based online photo-sharing system (figure 15), PhotoCorner, that allows home users to upload, manage, view, search, and share their photo albums anywhere from both web browsers as well as mobile wireless devices (e.g. PocketPC). The goal of such system is answer to the problem faced by home users that want to access their albums anytime, anywhere.

Images can be retrieved by several means, namely by text queries, by image examples and by sketching. The figure 15 presents on the right part a query based on sketching that expresses the fact that the user is looking for images containing people in the middle of the photograph, surrounded on the left and right-hand sides by foliages. The results are then presented in the center part of the figure. The right part of figure 15 is dedicated to the different functions proposed by the system, like access to previously defined albums, access to friends' albums, and creation of albums.

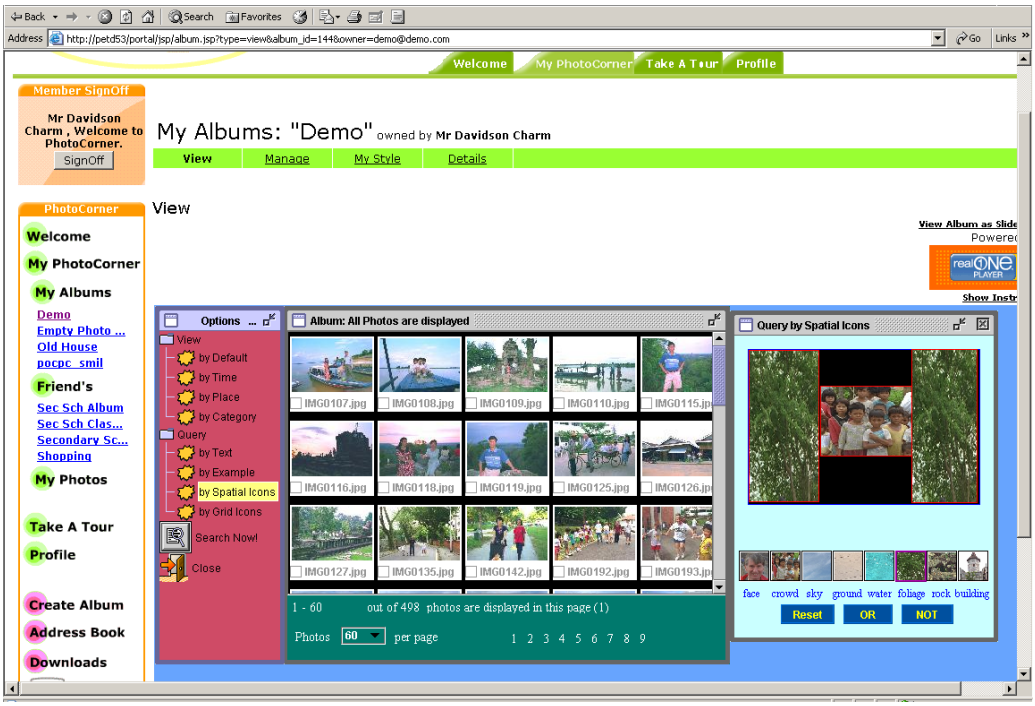

Figure 15. PhotoCorner Web-Based Interface.

The system also provides a user-friendly authoring tool for adventurous users to create their own slide show presentation styles for the photo albums, if they do not like the slide show templates provided. As the authoring tool is based on the open and declarative SMIL (Synchronized Multimedia Integration Language) W3C standard [SMIL 2001], the users can enjoy a truly multimedia viewing experience of their memorable photos synchronized with background music, narrations, text captions, and transitions, over the web (desktop and TV) and on mobile devices. The figure 16 presents the authoring tool used to create such animated presentations. Going from top to bottom, the user decides the style she/he wants for the presentation, the "Slide Layout" part allows the user to define when to present the photographs and what information to display along the photograph themselves, the "Displaying of Photos" lists the photographs to be presented, as well as the transition effects that can be used between photographs. Figure 17 presents the first two images from the presentation generated in figure 16.

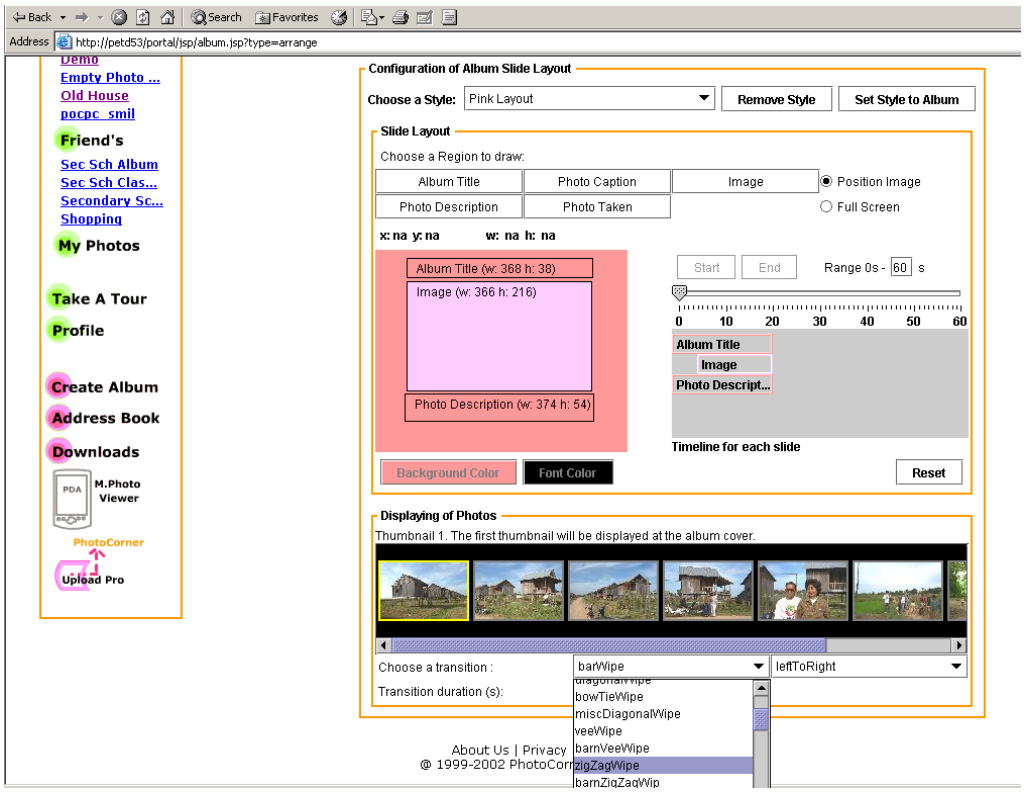

Figure 16. PhotoCorner Authoring Tool.

Figure 17. Samples from photographs presentation.

5. Experiments

This section is dedicated to present evaluation results from the retrieval functionalities of the prototype.

Our experiments were conducted on a set of 2400 real family photographs collected over a period of 5 years. Figure 18 displays typical photographs in this collection and Figure 19 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. The 2400 photographs were indexed automatically as described in Section 4.2. The detection of faces in the photographs was further enhanced with specialized face detector [Rowley et al. 1998]. The overall number of labels for VK and CG are 85 and 110 respectively. CG also uses 48 relations, and applies symmetry and transitivity among relations when needed.

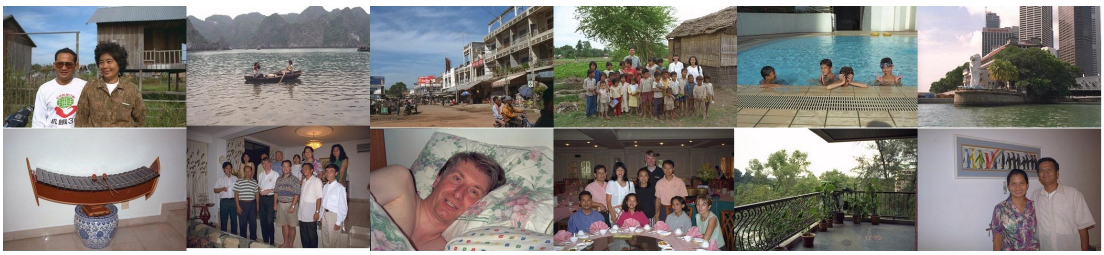

Figure 18. Typical family photographs used in our experiment.

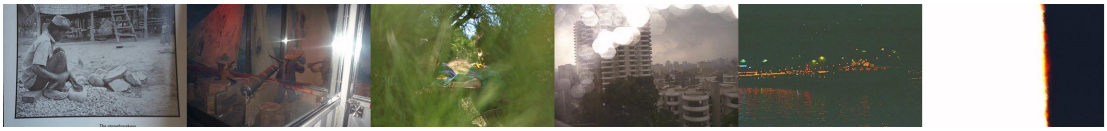

Figure 19. Some of the inferior quality family photographs.

Table 1. The queries defined on the family photograph collection.

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

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). For each query listed in Table 1, we selected 3 relevant photographs as QBE input to VK subsystem and constructed relevant textual query terms as QBS input to 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 focus initially on the general results related to the experiment. The figure 20 presents the average recall/precision results (over the 24 queries) obtained for the VK and CG only respectively, and for the best combination of both. 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 combined result shows that our integrated approach attains an increase of precision (+8.1% compared to the VK approach), implying that the integrated approach outperforms each of the individual subsystem. We also notice that the improvement of the precision values at the 0.2 and 0.4 recall points is greater than 11%, which leads us to indicate that the combination seems to favor the increase of precision for 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.

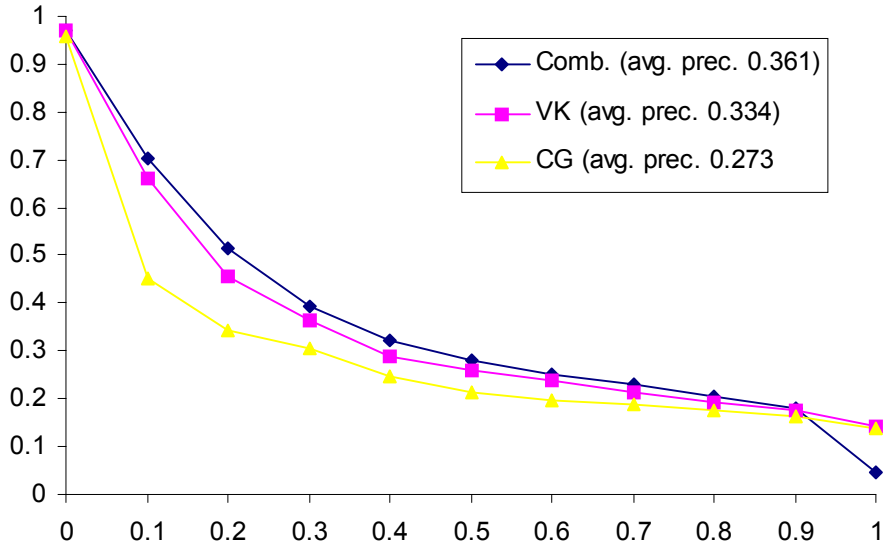

Figure 20. Recall vs Precision.

To be more precise about the results for low recall values, Table 2 presents the average precision values at 20, 30, 50, and 100 documents for the VK, CG and 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 2. Precision at 20, 30, 50, and 100 documents.

	VK	CG	Comb.
Avg. prec. at 20 doc.	0.592	0.408	0.621
Avg. prec. at 30 doc.	0.517	0.363	0.569
Avg. prec. at 50 doc.	0.437	0.330	0.488
Avg prec. at 100 doc.	0.351	0.305	0.403

If we analyse the results of Table 2 according to numbers of relevant documents, we notice that at 20 documents the recall value is increased by around 3%; so there is around 0.6 more document retrieved on 20 compared to the VK results. The same process applied to the 30, 50 and 100 documents precision values gives that the combination finds around 1.5 more relevant documents in the first 30, 2.5 for the first 50 documents and 5.21 more relevant documents in the first 100. Both results shows that the combination of VK and CG approaches that consider different levels of representations provides better results that worth the combined use of VK and CG approaches.

7. Conclusion

In this chapter, we have first analyzed the needs of digital home photo album users based on both the actual technological trends as well as by means of a user survey. This analysis brings out certain key points to be kept in mind for a designer of such digital albums:

- Need to build usable systems, specially given the enormous diversity of the users
- Need to share photos as well as store photos
- There will always be a plethora of data formats, capture and storage devices as well as communication channels
- Sharing can be on a global scale using yet should allow for crisp customization
- Albums must observe and learn (or be taught) quirks, preferences and nuances of its user

We have then described in detail about our efforts in the DIVA project to build innovative and functional tools that simplify the home users' effort in organizing, retrieving, viewing, and sharing their digital image albums. Since home photos are meant

for sharing, we need to locate individual images from a collection that conveys a particular thought or mood or idea. In order to do this, adequate content description (metadata) needs to be captured and stored alongside the image. Hence, the ability to annotate photos based on their semantic content is a vital and challenging task. We have thus primarily emphasized our work on some novel content-based semantic indexing and retrieval techniques. The promising experimental results for these techniques have also been outlined. We subsequently presented our attempts on home photographs management via our DIESKAU and PhotoCorner systems, especially on the retrieval and authoring aspects.

Given that home photo collections already represent an enormous amount of our civilizations' heritage by capturing silhouettes of millions of peoples' daily lives – representing snapshots of celebrations as well as their relationships, building suitable digital albums is an extremely important technological endeavor which needs concentrated attention. We therefore believe that striving to achieve these eminently worthwhile goals will yield rich and satisfying research opportunities for technological innovations in the fields of image analysis and vision processing, efficient and effective data storage and retrieval, and human computer interaction among others. While work has already been started as in the case of the DIVA project and the works presented in the references section, a lot more remains to be done before our full set of objectives are met. Our long-term collective desire is to enrich the home users' visual experience in creating and enjoying their digital memories, thereby allowing users to fully participate in this age of visual literacy.

References

[3D-Album 2002] Photo Album Software, http://www.3d-album.com

[Bach et al. 1996] J. R. Bach, C. Fuller, A. Gupta, A. Hampapur, B. Horowitz, R. Humphrey, R. Jain, and C. Fe Shu. The virage image search engine: An open framework for image management. In Proceedings SPIE Storage and Retrieval for Image and Video Databases, volume 2670, pp. 76-87, 1996.

[Chalfen 1998] R. M. Chalfen, Family Photograph Appreciation: Dynamics of Medium, Interpretation and Memory, Communication and Cognition 1998, 31(2-3), pp.161-78, 1998.

[Chen et al. 2002] J. Chen, T. Tan and P. Mulhem, SmartAlbum – Toward Unification of Approaches for Image Retrieval, International Conference on Pattern Recognition, Québec, August 2002, Vol. 3, pp. 983-986

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

[Faloustos et al. 1994] Faloutsos, C., Barber, R., Flickner, M., Hafner, J., Niblack, W., Petrovic, D., & Equitz, W., Efficient and Effective Querying by Image Content. Journal of Intelligent Information Systems, 3, 1994, pp. 231-262.

[FlipAlbum 2002] FlipAlbum, http://www.flipalbum.com

[Holt et al. 1997] Bonnie Holt and Ken Weiss. The QBIC Project in the Department of Art and Art History at UC Davis. In The 60th ASIS Annual Meeting 1997, Digital Collections: Implications for users, Funders, Developers and Maintainers, volume 34, pages 189-195. ASIS, 1997.

[Gargi et al. 2002] U. Gargi, Y. Deng, D. R. Tretter, Managing and Searching Personal Photo Collections, HP Laboratories Technical Report HPL-2002-67, Palo Alto, USA, 2002.

[Graham et al. 2002] A. Graham, H. Garcia-Molina, A. Paepcke amd T. Winograd, Time as Essence for Photo Browsing Through Personal Digital Libraries, International Workshop on Visual Interfaces for Digital Libraries, Portland, USA, 2002.

[Gupta 1995] A. Gupta, "Visual information retrieval: a Virage perspective," white paper, Virage Inc, 1995.

[Keng and Schneiderman 2000] H. Kang and B. Schneiderman, Visualization Methods for Personal Photo Collections: Browsing and Seraching in the PhotoFinder, ACM CIKM 2000, New York, USA, pp. 1539-1542, 2000.

[Kuchinski et al. 1999] A. Kuchinski, C. Pering, M. L. Creech, D. Freeze, B. Serra and J. Gwizdka, Fotofile: A Conusmer Multimedia Organization and Retrieval System, ACM CHI'99, Pittsburgh, USA, pp. 496-503, 1999.

[Leow 2002] W. K. Leow, The Algebra and Analysis of Adaptive-Binning Color Histograms. Tech. Report No. TRB8/02, Dept. of Computer Science, School of Computing, National University of Singapore, 2002.

[Leow and Li 2001] W. K. Leow and R. Li, Adaptive binning and dissimilarity measure for image retrieval and classification. In Proc. IEEE Conf. on Computer Vision and Pattern Recognition, 2001.

[Lim 2001] 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.

[Lu et al. 2000] Lu, Y., Hu, C., Zhu, X., Zhang, H. and Yang, Q. A Unified Framework for Semantics and Feature Based Relevance Feedback in Image Retrieval Systems, The 8th ACM Multimedia International Conference, November 2000, Los Angeles, CA, pp. 31-37.

[Luo and Etz 2002] J. Luo and S. Etz, A physics-motivated approach to detecting sky in photographs, International Conference on Pattern Recognition, Canada, Vol. I, pp. 155-158.

[Ma and Manjunath 1996] W. Y. Ma and B. S. Manjunath, Texture Features and Learning Similarity, In Proc. IEEE Conference on Computer Vision and Pattern Recognition, 425-430, 1996.

[Ma et al. 1991] W. Y. Ma and B. S. Manjunath, NETRA: A Toolbox for navigating large image databases, IEEE International Conference on Image Processing (ICIP'97), USA, Vol. I, pp. 568-571, 1997.

[Mechkour 1995] M. Mechkour, An Extended Model for Image Representation and Retrieval. Proceedings of the International Conference on Database

and Expert System Applications (DEXA'95), UK, pp. 395-404, 1995.

[Mills et al. 2000] T. J. Mills, D. Pye, D. Sinclair amd K. R. Wood, Shoebox: A Digital Photo Management System, AT&T Technical Report 2000.10, 2000.

[Minka and Picard 1997] T. P. Minka and R. W. Picard, Interactive learning using a "society of models", Special issue of Pattern Recognition on Image Databases, 30(4), 1997.

[MPEG-7 2001] MPEG-7 Committee, Overview of the MPRG-7 Standard (version 6.0), Report ISO/IEC JTC1/SC29/WG11 N4509, J. Martinez Editor, 2001.

[Mulhem et al. 2001] P. Mulhem, W. K. Leow, and Y. K. Lee, Fuzzy conceptual graph for matching images of natural scenes. In Proc. Int. Joint Conf. on Artificial Intelligence, 1397-1402, 2001.

[Mulhem and Lim 2002] P. Mulhem and J.H. Lim, Symbolic photograph content-based retrieval, To appear in *Proc. of ACM CIKM 2002, McLean, VA, USA, Nov. 4-9, 2002*.

[Ounis & Pasça 1998] I. Ounis and M. Pasça, RELIEF: Combining expressiveness and rapidity into a single system, ACM SIGIR 1998, Melbourne, Australia, pp. 266-274, 1998.

[Puzicha et al. 1999] J. Puzicha and J. M. Buhmann and Y. Rubner and C. Tomasi, Empirical Evaluation of Dissimilarity for Color and Texture, International Conference on Computer Vision 99 (ICCV'99), Kerkyra, Greece, pp. 1165-1172, 1999.

[Rodden 1999] K. Rodden, How People Oranise Their Photographs?, BCS IRSG 21st Annual Colloquium on Information Retrieval Research (BCS Electronic Workshops in Computing), Glasgow, UK, 1999.

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

[SMIL 2001] Synchronized Multimedia Integration Language 2.0, W3C Recommendation, August 2001, http://www.w3.org/TR/2001/REC-smil20-20010807 .

[Smith et al. 1996a] J. R. Smith and S. F. Chang, VisualSeek: A Fully automated content-based image query system, ACM Multimedia'96, Boston, USA, 1996, pp. 426-437.

[Smith et al. 1996b] J. R. 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.

[Schneiderman and Kang 2000] Ben Shneiderman, Hyunmo Kang (July 2000) Direct Annotation: A Drag-and-Drop Strategy for Labeling Photos, International Conference on Information Visualisation (IV2000). London, England, pp. 88-95, 2000.

[Srihari and Zhang 2000] R. K. Srihari and Z. Zhang, Show&Tell: A Semi-Automated Image Annotation System, IEEE Multimedia, 7(3), pp. 61-71, 2000.

[Salton and Buckley 1988] G. Salton and C. Buckley, Term-weighting approaches in automatic text retrieval, Information Processing and Management, Vol. 24, N. 5, pp. 5513-523, 1988.

[Sowa 1984] J. F. Sowa. Conceptual Structures: Information Processing in Mind and Machines. Addison-Wesley, Reading (MA), USA, 1984.

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

[Wenyin et al. 2000] L. Wenyin C. Hu and H. Zhang, iFind-a system for semantics and feature based image retrieval over Internet, The 8th ACM Multimedia International Conference, November 2000, Los Angeles, CA, pp. 477-478.