What's beyond query by example?

Nozha Boujemaa, Julien Fauqueur, Valérie Gouet

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

What's beyond query by example?

Nozha Boujemaa — Julien Fauqueur — Valérie Gouet

N° 5068

Décembre 2003

THÈME 3

ISSN 0249-6399 ISRN INRIA/RR--5068--FR+ENG

What's beyond query by example?

Nozha Boujemaa , Julien Fauqueur , Valérie Gouet

Thème 3 — Interaction homme-machine,
images, données, connaissances
Projet IMEDIA

Rapport de recherche n° 5068 — Décembre 2003 — 22 pages

Abstract: Over the last ten years, the crucial problem of information retrieval in multimedia documents has boosted research activities in the field of visual appearance indexing and retrieval by content. In the early research years, the concept of the "query by visual example" (QBVE) has been proposed and shown to be relevant for visual information retrieval. It is obvious that QBVE is not able to satisfy the multiple visual search usage requirements. In this paper, we focus on two major approaches that correspond to two different retrieval paradigms. First, we present the partial visual query that ignores the background of the images and allows a straight user expression on its visual interest without relevance feedback mechanism. The second retrieval paradigm consists in searching for the user mental image when no starting visual example is available. This new approach relies on the unpervised generation of a visual thesaurus from which query by logical composition of region categories can be performed. This query paradigm is closely related to that of text retrieval. Mental image search is a challenging and promising issue for retrieval by visual content in the forthcoming years since it allows different rich user expression and interaction modes with the search engine.

Key-words: query by visual example, partial queries, region of interest, points of interest, mental image, logical composition of region categories, visual thesaurus of regions, user expression

What's beyond query by example?

Résumé : Au cours de la dernière décennie, le problème fondamental de la recherche d'information dans les bases multimédia a alimenté les travaux dans le domaine de l'indexation et de la recherche par le contenu visuel. Initialement, le paradigme de "recherche par l'exemple visuel" a été proposé et s'est avéré pertinent pour la recherche d'information visuelle. Ce paradigme n'offre qu'une solution limitée aux divers besoins existants. Dans cet article, nous nous intéressons à deux approches correspondant à deux paradigmes distincts de recherche. Nous présentons, d'une part, un paradigme de requêtes partielles permettant d'ignorer le fond de l'image et qui permettant l'expression directe de parties d'image pertinentes sans mécanisme de bouclage de pertinence. Le second paradigme permet la recherche à partir d'une image mentale de la cible et ne nécessite pas d'image exemple initiale. Cette nouvelle approche repose sur la génération non-supervisée d'un thesaurus visuel à partir duquel peuvent être formulées des requêtes par composition logique de catégories de régions. Ce dernier paradigme est semblable à celui de la recherche de texte. La recherche par image mentale constitue un axe prometteur dans le domaine de la recherche par le contenu visuel dans la mesure où il offre des modes riches et différents d'interaction avec le moteur de recherche.

Mots-clés : recherche par l'exemple, requêtes partielles, régions d'intérêt, points d'intérêt, image mentale, composition logique de catégories de régions, thesaurus visuel de régions, expression utilisateur

1 Problem statement

The amount of available multimedia documents has steadily increased in later years and with it the need for efficient organization and retrieval of this information when needed. Simple arrangements of items and immediate lookup is no longer sufficient in a world more interested by the content than by the description tags found in most archives. These growing needs have boosted research activities in the field of content-based image retrieval (CBIR) that used to be achieved thanks to textual annotation.

In the last decade, the concept of "query by visual example" (QBVE) has been introduced [27][10] and has shown the relevance of the low-level-based visual information retrieval. This approach consists in sending the *entire* image (i.e. its computed low-level signatures) to the search engine as a visual query. Hence beside human-based metadata (textual annotation) that usually bring semantic information, these machine-based metadata related to the physical content (low level features) become available as an information retrieval support. All major institutions, in industry as well as in academic and public research groups, investigated this field.

In the case of QBVE (in the sense of full image query), retrieval results express an *overall global* visual similarity, thus an *approximate* similarity. In this context, we may have two images with different image components ("objects") with different shapes and appearances, but which remain globally similar. For some given visual queries, this leads to differences between user intention/target and retrieved results. Starting from these observations, our community has discussed the concept of "semantic gap" through different approaches. This was also the statement of the fact that the QBVE paradigm is not able to satisfy the multiple visual search requirements.

There are several ways to deal with the semantic gap. One prior work is to optimize the *fidelity* of physical-content descriptors (image signatures) to visual content appearance of the images. The objective of this preliminary step is to bridge what we call the "numerical gap". To minimize the numerical gap, we have to develop efficient images signatures (compact and visually consistent, eg. [7]). The weakness of visual retrieval results, due to the numerical gap, is often confusingly attributed to the semantic gap. We think that providing richer user-system interaction allows user expression on his preferences and focus on his semantic visual-content target.

Beyond QBVE, rich user expression comes in a variety of forms:

1. allow the user to notify his satisfaction (or not) on the system retrieval results method commonly called "relevance feedback". In this case, the user reaction expresses more generally a *subjective preference* and therefore can compensate for the semantic gap between visual appearance and user intention,

2. provide *precise* visual query *formulation* that allows the user to select precisely its region of interest and pull off image parts that are not representative of his visual target,

3. provide a mechanism to search for the user mental image when *no starting image example is available.* Part of this work will be published in [5].

Besides, combined image and text indexing and retrieval approaches are of great interest for the semantic gap reduction and are heavily investigated. However, this item is out of the scope of this paper.

The discussions of the first and the second item are closely related. However, let us linger over the following question: relevance feedback, what for? What are the objective and the use of this mechanism?

This mechanism was the initial way for user expression commonly investigated for semantic gap reduction. It was mostly used to *compensate for all mismatches* between user query and retrieval results. We strongly think that this mechanism should be reserved to subjective/semantic user preference or to concept search. For example, let us consider the search for the concept "Cézanne paintings" among an image database of masterpieces. For a given Cézanne painting query, the search engine could find as similar masterpieces which are not Cézanne's and conversely. By relevance feedback mechanism, the user helps the system to find out what visual appearance is important to reach [6]. Here the criterion is the user satisfaction.

We notice that the use of relevance feedback has sometimes overstepped its objectives and its abilities. One frequent example of mistaking is the following: among animals database, we would look for images that depicts "tiger in a forest", "tiger in a savannah", "tiger in a desert". The target here is to retrieve tiger images among the other animals independently from the nature of image background. For this family of queries, it is more appropriated to allow the user selecting the tiger image surface than to perform feedback on the entire image through several iterations to help the system to understand the user visual target. When the user points the object of interest with partial query, the background image signature is considered as noise.

The third item concerns the mental search paradigm. In this case, there is no starting image (or region) example available for the user. There is only a starting mental image from an event or a context memory. The objective is then: how could we build a search engine to reach this mental image? This question is also related to the "page zero" problem [9] (what images should be shown first to the user to formulate the query?). There is a non-unique approach to provide a solution. In one of the earliest approach ([21] and [15]), the system suggests two (or a list of) possible target images (entire images) to the user who points to the system what are the images that are more likely close to his mental image. The objective is to minimize the number of iterations to locate the target. We notice that in this case the system questions the user and not the contrary. A part from this kind of statistical model, there is another approach for mental image search.

In this paper, we present and discuss two local image signatures which provide local image description and allows partial querying. These local methods reduce explicitly the semantic gap as they integrate more appropriate user interaction with the search engine

within a precise query formulation. In the second part of the paper, we present a new visual retrieval paradigm based on target image search from visual thesaurus formed by image region categories. This paradigm allows the user to express his preference by logical composition of image parts over this visual thesaurus to reach a mental image when he has *no starting image examples.*

2 Partial visual selection for precise query

When the user is interested to retrieve only similar parts or objects of an image, a local selection must be considered. The idea consists first in localizing features of interest in the image and second in characterizing the primitives obtained with local descriptors. We could have weak requirements on partial retrieval precision for some applications and for others have hard constraints on the partial patch configuration such as photometric and/or geometric properties.

In this context, two main classes of primitives are relevant to characterize parts in an image: region segmentation and description for homogeneous area similarity search [11] and points of interest detection and description [16] when considering heterogeneous area. We will first present these two approaches and then discuss differences and complementarities between these two methods for partial query formulation.

2.1 Regions of Interest

We present here the approach we develop at the IMEDIA group to perform region-based visual query. The key idea, in this context, consists in considering rough image segmentation but in the same time a fine visual appearance description. Indeed, we state that we do not need high precision nor sub-pixel contour detection for image retrieval. Let us consider again the tiger visual retrieval example. It is sufficient to catch a piece of tiger texture surface to be able to retrieve images that contain tigers (or tiger-like surfaces). We do not need to segment precisely legs or other contour details. On the other hand, we compensate the roughness of this object extraction by fine adaptive region appearance description. We consider this choice as a reasonable compromise since an optimal and generic segmentation method is a utopia.

2.1.1 Region extraction

Detected regions should encompass a certain visual diversity to be visually characteristic, using a *coarse* segmentation. We want to stay beyond a too fine level of spatial and feature details. This coarseness makes regions a complementary approach to points of interest that rather characterize high spatial frequencies.

Figure 1: Example of segmented images. Small discarded regions are shown in gray. More examples of segmented images are available online [1].

The adopted segmentation approach, proposed in [11], is unsupervised and fast. It is based on the clustering of Local Distributions of Quantized Colors (referred to as *LDQC's*). Extracted in pixel neighborhoods this primitive captures the local color variability.

LDQC's are extracted as follows. For a more compact representation of colors in LDQC histograms, a color quantization of the image is first performed. Pixels in the Luv space are grouped using the Competitive Agglomeration clustering algorithm (referred to as *CA*, see [14]) which has the major advantage to automatically determine the number of clusters. We slide a window over pixels and evaluate the corresponding local distribution over the quantized color set. To each window (or neighborhood) corresponds a LDQC. All extracted LDQC distributions are clustered using CA and the Color Quadratic Form Distance [19]. Regions are defined as connected pixels in image space which have the same LDQC cluster tag. A Region Adjacency Graph is generated to merge or discard small regions, to improve spatial coherence of detected regions. We invite the reader to refer to [11] or [13] for more details on this segmentation scheme. Figure 1 illustrates some examples of our coarse segmentation scheme.

On a standard 2GHz PC, segmenting a 500x400 image takes 1.9 second on average. On an 11,479 image database, 5.2 regions per image are extracted on average.

This region extraction scheme is employed for both region-based retrieval schemes: retrieval by example region and retrieval by logical composition of region categories (see section 3).

2.1.2 Region description and retrieval

We suppose now that all images have been segmented. Given a region in an image, the user may wants to find other similar regions in the database regardless the background. Regions may correspond to salient areas such as a sky, a field, a road or may roughly correspond to an "object" such as a car, a face, an animal...

The problem consists in comparing the visual appearance of a query region to all regions in the database. The key idea in our approach detailed in [11] and [13] relies on the extraction of *coarse* regions, which are compared using a *fine* visual description, such that regions are specific against each other in the database. Existing color descriptor for regions as in VisualSeek [28], Blobworld [8] and Netra [22] are based on a fixed palette of

approximately 200 colors (between 166 and 256) to represent the entire color space. While such an approximate color representation is well suited to describe and retrieve images by their *global* content using traditional color histogram, regions are by construction more homogeneous and more numerous require a finer color representation to be distinguished from one another. In [11], the color variability region descriptor ADCS (Adaptive Distribution of Color Shades) is proposed. It is based on the distribution over a fine adaptive color binning of each region. A region index consists of the set of color shades specific to each region and their corresponding color populations. Color shades are determined for each region by a fine color quantization of the regions pixels using CA in the Luv color space. Any color from the full color space can be used as a color shade and does not depend on a fixed color palette. Compared to usual region color histogram, ADCS is more compact and more accurate to represent regions color variability.

Since color shades are specific to each region, measuring similarity between two ADCS requires an adapted distribution distance. The Generalized Form of Color Quadratic Distance [13] provides an efficient way to compare two color distributions *whatever their respective color binning.*

Depending on the type of searched regions, additional geometrical features may be combined with ADCS. For instance, a query initiated by an example of snow region may return regions of pale sky if they have similar gray/white distributions. By constraining the position of searched region in the lower part of the image, one can retrieve snow regions more efficiently. On the other hand to search regions of vegetation, position and area may not be particularly discriminant. In our Ikona platform [4], we allow the user to interactively set relative weights between ADCS (see [11]), position and area in the region comparison process depending on the type of searched regions.

Figure 2: Retrieval results from top-left lavender region. Retrieved regions are outlined in white. Color (ADCS descriptor), area and position are the region descriptors involved in the search. Other regions of lavender are retrieved although having various textures.

Figure 2 illustrates a query on a lavender region using the combination of ADCS, area and position. Retrieved regions are similar to the query region with respect to both photometric and geometric features.

Indexing all regions in a 400x500 image takes 0.8 second on average. Retrieving regions similar to a given example region from a photostock database of 11,479 images with 56,374 regions, takes 0.8 second at most on a 2GHz PC.

2.2 Points of Interest

In this section, we present an alternate method to perform partial query with harder requirements on the retrieved patch visual properties (photometric or geometric). This section is structured as follows: section 2.2.1 describes the color local descriptor used for sub-image or object retrieval which is based on points of interest. In section 2.2.2 we present our point-based image retrieval system according to this local description. Finally two practical scenarios are presented at section 2.2.3.

2.2.1 A local color image description

Figure 3: Example of Harris Color Point extraction.

When applied to image retrieval, image matching based on points of interest needs points with excellent *repeatability*, i.e. points that can be extracted from images with the same accuracy and under various conditions like viewpoint or illumination changes. Many point extractors exist for gray value images, for example [20] [23] and only one for color images [24]. It has been demonstrated [18] that it is the color operator which fits better for the required repeatability. We use this one to extract the points in the whole image during the indexing step.

In a second step, it is necessary to describe the points in a feature space, which is function of the photometric information around the point. Some approaches exist for grey value images [3] [25] [26]. For color images, the HCP solution (Harris Color Points) proposed in [16] consists in a characterization based on the combination of the Hilbert's differential invariants, which is invariant to image translation and rotation, robust to scale change (with a multi-scale approach), to illumination changes (if images or features are locally normalized), and robust to image coding and compression by considering low orders [17]. At first order for RGB images, we obtain the following features at point $\overrightarrow{x}$:

$$\overrightarrow{v}(\overrightarrow{x},\sigma) = (R \quad ||\nabla R||^2 \quad G \quad ||\nabla G||^2 \quad B \quad ||\nabla B||^2 \quad \nabla R.\nabla G \quad \nabla R.\nabla B)^T \qquad (1)$$

where σ represents the size of the Gaussian smoothing applied during the derivatives computation.

The similarity measure employed is the Mahalanobis distance δ^2, which takes into account the different magnitudes of the components and includes a model of noise. Such a feature space will be noted (V, δ^2) in the paper; its size depends on the order considered for the invariants computation.

2.2.2 Retrieval strategy

We present in this section the strategy we have adopted to retrieve the most similar images to the query one consisting of a voting algorithm.

If we consider the local descriptors presented above, an image is represented by a set of n points $\{p_i\}$ characterized in a feature space (V, δ^2). In this context, building an index for a database $\{I_j\}$ of N images consists in computing a set of $n \times N$ descriptors in (V, δ^2). Searching for an image or part of an image in the indexed image database comes to find in the space the closest points to the query points.

Let $\{q_i\}$ be the set of query points. The $\{r_i\}$ closest points to the $\{q_i\}$ query ones are characterized by scores which are function of the distance between the couples (r_i, q_i). A vote is computed for each image by considering a combination of the scores related to the matches (r_i, q_i) involved in the image. The most similar images to the query are characterized by the best votes. The complexity of the query can be efficiently reduced by organizing the indexes according to multidimensional structures.

In addition, a semi-local geometric characterization that considers the spatial relations between neighbor points of the same image can be added to enrich the photometric description [24].

2.2.3 Example of retrieval

We present in this section two typical scenarios of use of the local image description presented above.

Pattern retrieval: Digital art images are becoming widely available for cultural heritage material sharing. Moreover, one can find art resources across the web with online browsing facilities (e.g. online museum collections, companies' websites[1] which propose online galleries of antiques to sell, etc.). In the best cases, the user can interrogate the database by using keywords.

Other websites[2][3] , dedicated to stolen works of art, make an inventory of stolen items after registration to police which although sometimes associate images, do not allow any kind of visual interrogation.

We present in this section one practical interrogation scenario on such databases, according to the content-based indexing and retrieval approaches presented in the previous section. The database used for the experiments contains 1077 color images of antiques[4] with different viewpoints, illumination conditions and partial occlusions.

Let us suppose that a collector got a vase (see left picture) and that he would like to acquire other items from the same collection. If he does not know that this piece is a Greco-Roman porcelain urn characterized by scrolling acanthus leaves, it will not be easy for him to exploit keyword-based search tools. On the other hand, the local content-based approach presented will be able to perform this task in a better way, as illustrated in figure 4. This result presents the interface of our CBIR system Ikona [4], here implemented with the HCP descriptor. The first image on the upper-left shows the query area defined manually by the user. Only the points of interest contained in this window are used during the retrieval step. The retrieval results are then presented by

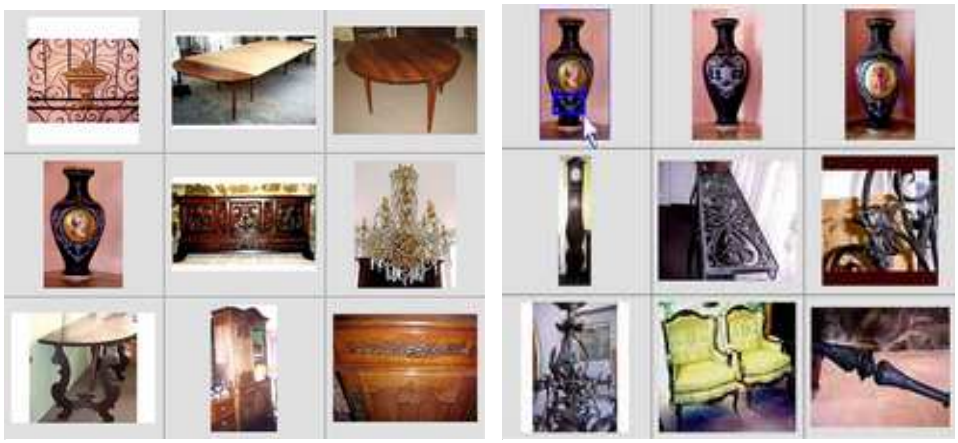

Figure 4: An example of pattern retrieval: interactive query selection by the user represented by the blue rectangle area in top-left image. Random view of antiques image database content (left). The partial query results by decreasing similarity value (right).

decreasing order of vote. We see that the database effectively contains two other objects characterized by these particular blue leaves and belonging to the query series.

Background retrieval in different contexts: In the scenario of figure 5 below, we want to retrieve the images of the TV series which contain a particular background selected by the user. Indeed, we focused on the upper left part of the image, which shows partially something like a wine storeroom.

The retrieval was performed on this particular region, described by about thirty interest points, as exhibited on the second image of the figure. The matching results were sorted by decreasing order of the image scores obtained and the best ones are presented on the second row of the figure. We show that the query area has been retrieved in five images, which involve the same room but with different characters. The resulting images differ from the image query in global shape and color, in viewpoint and present some occultation. Global indexing approaches naturally would not have given interesting results for this class of query. Approaches based on region segmentation would not allow the user to make the query on this part, since it represents small regions not easily detectable.

The scenario presented here is currently used by the French Judicial Police as an investigation aid with image similarity retrieval. Other examples of sub-image retrieval with points of interest can be seen at [2].

[1] http://www.faccents.com/.

[2] http://www.gazette-drouot.com/vols.html

[3] http://www.gendarmerie.defense.gouv.fr/judiciaire/

[4] Images presented in this section are provided by the "French Accents" company (http://www.faccents.com/).

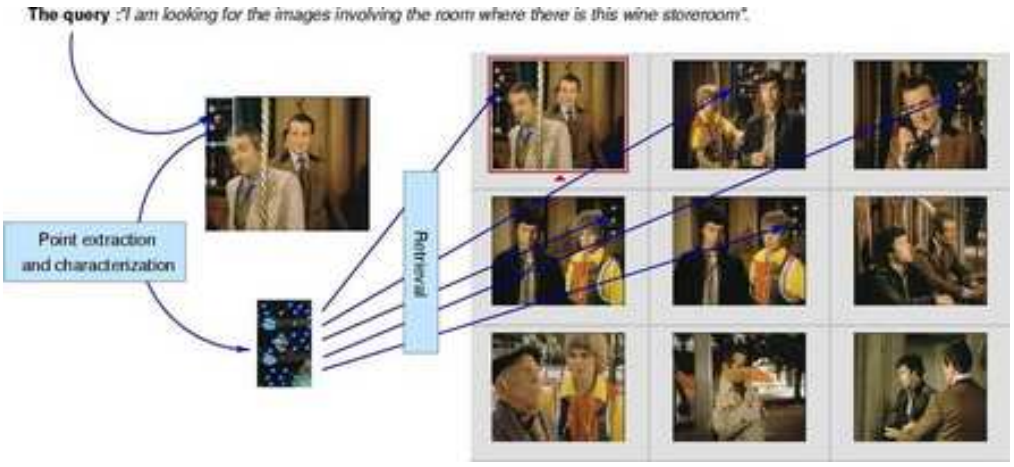

Figure 5: A scenario of background retrieval with different foreground visual information.

2.3 Discussion of local image signatures typology

In sections 2.1 and 2.2 we proposed two approaches to characterize and retrieve images parts: regions and points of interest. Both allow the user to select an image area and retrieve images which contain a similar area. But they really differ in terms of visual representation and usage and their complementarity is an asset to satisfy different partial query scenarios.

In the first case, a region is defined as an area of connected pixels which have similar local description (the LDQC) in an off-line and unsupervised phase. Regions are coarse and correspond to dominant and homogeneous areas in images. Their description takes into account all pixels of the region by measuring their color distribution (with ADCS): this is a statistical description, i.e. the smaller and the fewer the details are and the lower their participation in the region description is. The statistical nature of the description allows an approximate search based on the overall appearance of detected regions.

In the second case, points are detected with the HCP detector and exist only on very small sites (a few pixels wide) which present a high photometric variability. As a consequence, uniform image sites are ignored. These points are then described by Hilbert differential invariants which characterize this photometric variability. User selection encompasses the set of detected points which are the query points. The voting scheme allows matching them with the best candidate points in each image. As a consequence, areas of interest are *interactively* defined. The pointwise matching scheme and the flexibility in the area definition make the point retrieval computationally expensive.

In figure 6, on the same three images regions and points were detected. We notice that in *uniform* (sky, snow) and *smooth* parts (green blurred background) no points are detected while segmentation easily detected these parts. On the other hand, points appear on large textured parts (grape stack, vegetation) and on details (wine bottles).

Figure 6: Local features: region segmentation and point of interest extraction on a particular scene. From left to right: original images, segmented images, images with region boundaries superimposed on originals, extracted points of interest. While points of interest detect image sites of high spatial frequency, regions are expected to detect large areas which are homogeneous with respect to a photometric primitive.

The two objective differences between both approaches are the advantage of flexibility for the area definition in the point approach and a much lower retrieval computational cost for the region approach.

Other differences depend on the nature of searched areas. If user's focus in images is large and homogeneous areas, region approach should be preferred since detected points will be non-existent or will correspond to noise. Conversely to search very small areas with characteristic details, region approach is irrelevant since segmentation may not have detected such small areas and even if detected, their description won't be relevant (see wine cellar scenario). Typically, a precise (i.e. with a fine level of visual details) search on salient details should be carried out with points. As a consequence, the usage of each approach should be motivated by the user target precision requirements. Otherwise, if the nature of searched areas is not clearly defined in advance, region approach should be used because of its faster response. Point-based retrieval is more precise but much more costly in the same time. Our ongoing research work is concerned with scalability of this approach.

3 Mental image search by regions composition from visual thesaurus

Be they global or partial based, existing CBIR systems all require a starting image or region example to perform a search. This approach is well suited to perform visual comparison between a given example and entries in the database, i.e. to answer to such a query: "find images/regions in the database similar to this image/region". But very often, the user doesn't have an example image to start the search. **When the target image is only a mental image in the user's mind**, the prior search of an example to perform the actual query by example is tedious, especially in the case of a multiple region query.

A new framework presented in [12] differs completely from this paradigm on both query and retrieval processes. Images are retrieved by *logical composition of region categories*. Categories of similar regions are generated by an off-line unsupervised process. A "photometric region thesaurus" of the database is derived from these categories. From this thesaurus in the query interface, the user can select the types of regions which should and should not appear in mental target image. It allows retrieving images very quickly from logical queries as complex as: "find images composed of regions of these types and no regions of those types".

3.1 Categorization and range query in the regions feature space

The database structure is based on the following principle. Images are first segmented by classification of LDQC, as proposed in section 2.1. All extracted regions in the database are indexed by a visual descriptor. We define the **region categories** (denoted $C_1, ..., C_P$) as the clusters of regions which have similar visual features. They are the basis of the definition of similar regions in the retrieval phase. Here we choose to characterize regions with their

mean color such that regions from the same category have similar mean color. It is important to note that other visual cues could be used such as color distribution, texture, position, area and some specific descriptor. For instance if a texture descriptor were used instead of mean color, each category would be expected to group regions of similar texture. Despite the straightforwardness of mean color description, we will see it is sufficient to form generic categories. Regions mean colors are determined in the Luv space, which is chosen for its perceptual uniformity. We cannot make a priori assumption concerning the well-definition of clusters of regions for any database. But what can always be guaranteed is an intra-category visual coherence by setting a fine clustering granularity.

Region categories are formed by grouping the regions descriptors with CA and a fine granularity. For each region category, its **representative region** is defined as the closest region to its prototype. Representative regions are used to identify each category in the query interface. Since similarity between regions will be defined, at a first level, as members of the same category, a fine clustering granularity will ensure the retrieval of very similar regions (hence high retrieval precision). At a second level, we also consider as similar regions which are in close categories (called "neighbor categories") to also allow high recall. This key idea allows achieving **range queries** in the regions feature space. The **neighbor category** of a category C_q of prototype p_q is defined as a category C_j whose prototype p_j satisfies: $|| p_q - p_j ||_{L^2} < \gamma$, for a given range radius threshold γ (which is adjusted at the query phase).

We call $N^\gamma(Cq)$ the set of neighbor categories of a category C_q. See figure 7 for an illustration of the definition of neighbors using the radius. Note that thanks to this range query scheme, the search is less dependent on the database partition into categories since all close categories are considered together.

The combination of homogeneous region categories with the integration of neighbor categories is the key choice in the definition of the range query scheme.

3.2 Image Retrieval by Composition

From this point on, regions aren't considered individually anymore but are totally identified to the category they belong to. With the help of all categories representative regions (see figure 3.2), the user will select Positive Query Categories (referred to as **PQC's**) and Negative Query Categories (**NQC's**). The PQCs correspond to the user-selected categories of regions which should appear in retrieved images. They are denoted as $\{C_{PQ_1}, ..., C_{PQ_M}\}$. The NQCs correspond to the user-selected categories of regions which should *not* appear in retrieved images and are denoted as $\{C_{NQ_1}, ..., C_{NQ_R}\}$. In its most complex form, a **query composition** is defined as the formulation: "find images composed of regions in these PQCs and no region from those NQCs". It is expressed as the list of PQC labels $\{pq_1, ..., pq_M\}$ and NQC labels $\{nq_1, ..., nq_R\}$.

Performing a query composition first requires to retrieve images which contain a region from a single PQC category denoted C_{pq} say. For a given category C_{pq} , we define $IC(C_{pq})$ to be the set of images containing at least one region belonging to category C_{pq}. To expand this search to a range query, we also take into account neighbor categories of C_{pq} by defining

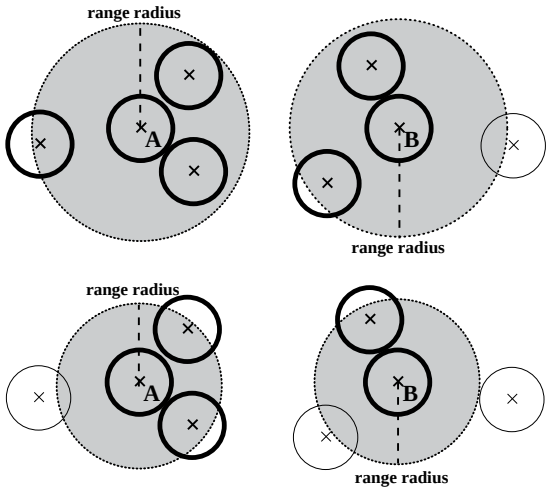

Figure 7: Range radius and neighbor categories: A and B are two categories. Neighbor categories are drawn with thick contours. Prototypes are identified by crosses. The gray disks cover the neighbor categories. A high radius (top) or a lower radius (bottom) integrates more or less neighbor categories to define the type of searched regions.

relevant images as those which have a region from category C_{pq} or from any of its neighbors:

$$\bigcup_{C \in N^{\gamma}(C_{pq})} IC(C) \tag{2}$$

Range radius threshold γ is set in the user interface.

To extend the query to all M PQCs: $C_{pq_1}, ..., C_{pq_M}$, we search images which have a region in C_{pq_1} or its neighbors and ... and a region in C_{pq_M} or its neighbors. The set S_Q of images satisfying this multiple query is now written as:

$$S_Q = \bigcap_{i=1}^{M} \Big[\bigcup_{C \in N^{\gamma}(C_{pq_i})} IC(C) \Big] \tag{3}$$

Then, to also satisfy the negative query we must determine images which contain a region from any of the R NQCs : $C_{nq_1}, ..., C_{nq_R}$. As before, neighbor categories should also be taken into account. So the set S_{NQ} of images containing the NQCs is written as:

$$S_{NQ} = \bigcap_{i=1}^{R} \Big[\bigcup_{C \in N^{\gamma}(C_{nq_i})} IC(C) \Big] \tag{4}$$

The set S_{result} of retrieved images which have regions in the different PQCs and which don't have regions in the NQCs is expressed as the set subtraction of S_Q and S_{NQ} :

$$S_{result} = S_Q \setminus S_{NQ} \tag{5}$$

This set S_{result} constitutes the final set of relevant images.

Unions, intersections and subtractions in the expression of S_{result} are directly equivalent to formulate the query with logical operators as illustrated in figure 10: `or` between the neighbors (in expressions of S_Q and S_{NQ}), `and` between query categories (also in S_Q and S_{NQ}), `andnot` for negative query categories (in expressions of S_{result}).

To evaluate the expression of S_{result} , the brute force approach would consist in testing, for each image in the database, if it contains regions belonging to the PQCs (and their neighbors) but contains no region from the NQCs (and their neighbors). Instead, to reduce dramatically this number of tests in a simple way, we use the fact that S_{result} is expressed as intersections and subtractions of image sets. The idea is to initialize S_{result} with one of the image sets and then discard images which don't belong to the other image sets. This initialization avoids testing individually each image of the database. We directly start off with a set of potentially relevant images. S_{result} will be gradually reduced as follows:

1. initialize S_{result} as the set $\bigcup_{N^\gamma(C_{pq_1})} IC(C)$.
2. discard images in S_{result} which do not belong to any of the other union categories (i=2,...,M) to obtain the intersections of S_Q . At this point, we have $S_{result} = S_Q$.
3. to perform the subtraction of S_{NQ} from S_{result} , discard in S_{result} images which belong to the negative-query union categories (i=1,...,R). We get $S_{result} = S_Q \setminus S_{NQ}$.

Gradually, S_{result} is reduced from $\bigcup_{N^\gamma(C_{pq_1})} IC(C)$ to $S_Q \setminus S_{NQ}$. By this approach, we'll see in next section that a significant fraction of the database is not accessed at all.

This retrieval scheme is easily implemented using three tables of association which provide associations between categories, neighbor categories and images. It is important to note that at retrieval time we don't deal with regions themselves but only with images and labels of region categories, so that we don't have to individually access the large number of regions in the database. Search process is very fast since it only involves elementary operations on integers, unlike classic search approaches which require distance computations between multidimensional feature vectors.

3.3 Results and query interaction

On a database of 9,995 images from Corel Photostock[5], 50,220 regions were extracted by segmentation. From these regions, 91 categories were generated by grouping their mean colors.

[5]Corel Photostock: http://www.corel.com

In the query interface (see figure 8), each category of regions is represented by its representative region which is defined as the region whose index is the closest to the category prototypes. Figure 8 shows a part of the 91 representative regions. The set of category representatives provides an overview of the types of regions. This set defines the **photometric region thesaurus** of the database. Any region category can be selected by its representative region to indicate that this type of regions should or should not appear in mental target image. Given the set of selected categories (see figure 9 the query is translated by the system as a logical expression of composition of region categories (see figure 10).

Figure 8: In the query interface, the user can select each of the 91 category representatives as a PQC or a NQC. A tick in a green box indicates that this type of region should appear in the mental target image and a tick in a red box indicates the type should not appear. This interface constitutes the "photometric region thesaurus".

Figure 9: A query example: the user wants images with a blue region and a gray region but with no green region. Range radius γ is set to 15.

Figure 10: Query of figure 9) is translated by the system as a logical expression of composition of region categories with the neighbor categories corresponding to each query category.

Consider an example of query composition. In the Corel database, to search cityscapes, the user will search images with a building, some sky and no vegetation. This can be translated into the following query composition: "images composed of a gray region and a blue region but no green region". Figure 9 and figure 10 illustrate this query and figure 11 (cropped screenshot) shows the relevant set of images retrieved for this query among the 9,995. In these images, gray regions match buildings, monuments or rocks and blue regions essentially match sky. The system also shows images which were rejected due to presence of a green region (see figure 12). It is interesting to notice that, on this query for cityscapes, rejected images correspond to landscapes (see figure 12)). We observed that *visual semantics* arises from the logical composition expressed by the user.

Images retrieved with such a simple system do satisfy the constraints of region composition. Its performance relies on the ability of the segmentation scheme to correctly detect salient regions and the nature of descriptor (from which regions are grouped) which should be chosen with respect to the

domain of application. Note that category and range radius selection can help the user improve interactively the retrieval performance depending on his satisfaction.

Retrieval scheme solely relies on accesses to three tables of association and no numerical feature distance is involved at this step (only at off-line grouping phase). Retrieval scheme is hence very fast: 0.03 second for most complex queries on a standard 498MHz PC. Besides, note that search in the database is neither exhaustive in terms of images, nor region categories, nor regions because database organization provides direct access to set of potentially relevant regions.

Figure 11: Retrieved images from the "cityscape" query: "gray region and blue region and not green regions". Images do contain a blue and gray region but no green region.

Figure 12: Images rejected from the "cityscape" query due to the presence of a green region. Rejected images turn out to depict *natural landscapes*.

When viewing categories as labels of similar regions, which are the constituting units of the database images, this indexing can be considered as *symbolic* rather than numerical. The following comparison with text retrieval can be made :

- image → document
- region → term
- region category → concept
- neighbor categories → similar/synonymous concepts
- set of region categories → thesaurus
- Query by logical composition → Google-like query [6]

3.4 Summary of the new paradigm

This new paradigm allows performing logical composition of region categories. The system allows retrieving images by query composition like: "find images with regions of theses types and not like those types". The originality of this approach relies on the grouping of similar regions into categories and has the following advantages:

- no required starting example region

[6] http://www.google.com

- query by image composition using regions categories
- natural region range query by interactive definition of neighbors categories
- efficient indexing and very fast image retrieval

Although a very simple color region feature is used, the constraint of composition in retrieved images seems to express some underlying "visual semantics" in images.

This framework is very simple and general. It can lead to further developments such as proposing a more perceptual arrangement of categories in query interface, integrating other region descriptors, handling spatial layout of the categories, developing a hierarchical region categorization to handle very large databases, association between visual categories prototypes and textual ontology to allow a visual-semantic search.

4 Conclusion

The most classical visual query paradigm QBVE, where the visual query concerns the whole image was useful to make the proof of feasibility for information retrieval by visual content. Once it is viable, we can fulfill the everyday user requirements beyond the QBVE simple query formulation.

We have presented some new user visual query formulation mechanisms. In all cases, these mechanisms allow the user to precise visual target selection, expression of visual target image composition including his preferences, with logical composition. The latest case concerns visual information retrieval when no starting image example is available but a mental image in the user mind. The main goal is to fit more and more to the semantic target of the user. Hence, the user is more and more interacting with the system by the means of new query paradigms.

Ongoing work, deal with scalability issues (particularly for point-of-interest description), different local image signatures with regions (handling spatial layout) and point features combination. Also, hybrid text and image indexing are heavily investigated to help bridging the semantic gap, particularly making the connection between textual ontology with the regions prototypes in the visual thesaurus.

References

[1] `http://www-rocq.inria.fr/~fauqueur/ADCS/`.

[2] `http://www-rocq.inria.fr/~gouet/HCP/`.

[3] A. Baumberg. Reliable feature matching across widely separated views. *IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 774–781, 2000.

[4] N. Boujemaa, J. Fauqueur, M. Ferecatu, F. Fleuret, V. Gouet, B. Le Saux, and H. Sahbi. Ikona: Interactive generic and specific image retrieval. *International workshop on Multimedia Content-Based Indexing and Retrieval (MMCBIR'2001), Rocquencourt, France*, pages 25–28, 2001.

[5] N. Boujemaa, J. Fauqueur, and V. Gouet. *What's beyond query by example? to appear in* Trends and Advances in Content-Based Image and Video Retrieval, *L. Shapiro, H.P. Kriegel, R. Veltkamp (ed.).* LNCS, Springer Verlag, 2004.

[6] N. Boujemaa, V. Gouet, and M. Ferecatu. Approximate search vs. precise search by visual content in cultural heritage image databases. *Invited paper in MIR workshop in conjunction with ACM Multimedia, Juan-Les-Pins, France*, 2002.

[7] N. Boujemaa C. Vertan. Upgrading color distributions for image retrieval: can we do better? *Proc. of International Conference on Visual Information System (VIS'00)*, pages 178–188, 2-4 Nov. 2000.

[8] C. Carson and al. Blobworld: A system for region-based image indexing and retrieval. *Proc. of International Conference on Visual Information System, LNCS vol. 1614*, pages 509–517, 1999.

[9] M. La Cascia, S. Sethi, and S. Sclaroff. Combining textual and visual cues for content-based image retrieval on the world wide web. *IEEE Workshop on Content-based Access of Image and Video Libraries (CBAIVL'98)*, june 1998.

[10] A. DelBimbo. *Visual Information Retrieval.* Morgan Kauffman, San Francisco, CA, 1999.

[11] J. Fauqueur and N. Boujemaa. Region-based retrieval: Coarse segmentation with fine signature. *IEEE International Conference on Image Processing (ICIP)*, pages 609–612, 2002.

[12] J. Fauqueur and N. Boujemaa. New image retrieval paradigm: logical composition of region categories. *IEEE International Conference on Image Processing (ICIP)*, pages 601–604, 2003.

[13] J. Fauqueur and N. Boujemaa. Region-based image retrieval: Fast coarse segmentation and fine color description. *Journal of Visual Languages and Computing (JVLC), special issue on Visual Information Systems*, 15(1):69–95, 2004.

[14] H. Frigui and R. Krishnapuram. Clustering by competitive agglomeration. *Pattern Recognition*, 30(7):1109–1119, 1997.

[15] D. Geman and R. Moquet. A stochastic model for image retrieval. *Congrès Francophone de Reconnaissance des Formes et Intelligence Artificielle (RFIA), Paris*, pages 173–180, 2000.

[16] V. Gouet and N. Boujemaa. Object-based queries using color points of interest. *IEEE Workshop on Content-Based Access of Image and Video Libraries (CBAIVL'01)*, 2001.

[17] V. Gouet and N. Boujemaa. On the robustness of color points of interest for image retrieval. *IEEE International Conference on Image Processing (ICIP), Rochester, USA*, 2002.

[18] V. Gouet, P. Montesinos, R. Deriche, and D. Pelé. Evaluation de détecteurs de points d'intérêt pour la couleur. *Congrès Francophone de Reconnaissance des Formes et Intelligence Artificielle (RFIA), Paris*, 2000.

[19] J. Hafner, H. Sawhnay, and al. Efficient color histogram indexing for quadratic form distance functions. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 17(7):729–736, July 1995.

[20] C. Harris and M. Stephens. A combined corner and edge detector. *Alvey Vision Conference*, pages 147–151, 1988.

[21] Thomas P. Minka Ingemar J. Cox, Matt L. Miller. The bayesian image retrieval system, pichunter: Theory, implementation and psychological experiments. *IEEE Transactions on Image Processing*, 9(1):20–37, 2000.

[22] W. Y. Ma and B. S. Manjunath. Netra: A toolbox for navigating large image databases. *Multimedia Systems*, 7(3):184–198, 1999.

[23] K. Mikolajczyk and C. Schmid. Indexing based on scale invariant interest points. *International Conference on Computer Vision (ICCV)*, pages 525–531, 2001.

[24] P. Montesinos, V. Gouet, and R. Deriche. Differential invariants for color images. In *Proceedings of* 14^{th} *International Conference on Pattern Recognition (ICPR'98)*, Brisbane, Australia, 1998.

[25] F. Schaffalitzky and A. Zisserman. Multi-view matching for unordered image sets. *European Conference on Computer Vision (ECCV)*, pages 414–431, 2002.

[26] C. Schmid and R. Mohr. Local grayvalue invariants for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 19(5):530–535, 1997.

[27] A. Smeulders, M. Worring, and S. Santini. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 22(12), 2000.

[28] J. R. Smith and S. F. Chang. Visualseek: A fully automated content-based image query system. *ACM Multimedia Conference, Boston, MA, USA*, pages 87–98, 1997.

Contents

Unité de recherche INRIA Rocquencourt
Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)

Unité de recherche INRIA Futurs : Parc Club Orsay Université - ZAC des Vignes
4, rue Jacques Monod - 91893 ORSAY Cedex (France)
Unité de recherche INRIA Lorraine : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)
Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)
Unité de recherche INRIA Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier (France)
Unité de recherche INRIA Sophia Antipolis : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399