Language-based Querying of Image Collections on the basis of an Extensible Ontology

Christopher Town
University of Cambridge, Computer Laboratory
15 JJ Thomson Avenue
Cambridge UK

David Sinclair
Waimara Ltd
115 Ditton Walk
Cambridge UK

Abstract

The design of a specialised query language for content based image retrieval (CBIR) provides a means of addressing many of the problems associated with commonly used query paradigms such as query-by-example and query-by-sketch. By basing such a language on an extensible ontology which encompasses both high-level and low-level image properties and relations, one can go a long way towards bridging the semantic gap between user models of saliency and relevance and those employed by a retrieval system.

This paper discusses these issues and illustrates the design and use of an ontological retrieval language through the example of the OQUEL query language. The retrieval process takes place entirely within the ontological domain defined by the syntax and semantics of the user query. Since the system does not rely on the pre-annotation of images with sentences in the language, the format of text queries is highly flexible. The language is also extensible to allow for the definition of higher level terms such as "cars", "people", etc. on the basis of existing language constructs through the use of Bayesian inference networks. The matching process utilises automatically extracted image segmentation and classification information and can incorporate any other feature extraction mechanisms or contextual knowledge available at processing time to satisfy a given user request.

Keywords

Image retrieval, query languages, ontologies, object recognition, language parsing

1 Introduction

Query mechanisms play a vital role in bridging the *semantic gap* [17] between users and retrieval systems. There has however been relatively little recent work in addressing this issue in the context of content based image retrieval (CBIR). Most of the query interfaces implemented in current systems fall into a small group of approaches. In order to overcome the weaknesses of these methodologies, efforts have focussed on techniques such as *relevance feedback* ([52], [9], [41]) as a means of improving the composition and performance of a user query in light of an initial assessment of retrieval results. While this approach and other methods for improving the utility of user queries by means such as query expansion and by combining multiple query modalities have shown some promise, they do so at the risk of increased user effort and lack of transparency in the retrieval process.

This paper presents the notion of an ontological query language as a powerful and flexible means of providing an integrated query and retrieval framework which addresses the problem of the *semantic gap* between user and system. Further background information and a general motivation for such query languages is given in section 2 while section 4 introduces the *OQUEL* language as a concrete example for retrieval from photographic image collections. Section 3 discusses the basic language design and structure. In section 5 the process of query interpretation and retrieval is described further. The discussion is based on an implementation of the language for the *ICON* content-based image retrieval system. Those content extraction and representation facilities of ICON relevant to the present discussion are outlined in section 4.1. Section 6 gives quantitative performance results of OQUEL queries compared to other query modalities in the ICON system. The paper concludes with a summary in section 7.2 which also provides an outlook of further developments such as the potential role of natural language processing and possible extensions of the ontological language to video.

1.1 CBIR Query Mechanisms and the Retrieval Process

As has been noted elsewhere (e.g. [45]), research in content based image retrieval has in the past suffered from too much emphasis being placed on a system view of the retrieval process in terms of image processing, feature extraction, content representation, data storage, matching, etc.. It has proven fruitful in the design of image retrieval systems to also consider the view of a user confronted with the task of expressing his or her retrieval requirements in order to get the desired results with the least amount of effort. While issues such as the visualisation of query results and facilities for relevance feedback and refinement are clearly important, this paper is primarily concerned with the mechanisms through which users express their queries.

Adopting a user perspective, one can summarise most of the query methods traditionally employed by CBIR systems (see [45] and [40] for further references) and highlight their drawbacks as follows:

- *Query-by-example*: ([48], [9], [24]) Finding suitable example images can be a challenge and may require the user to manually search for such images before being able to query the automated system. Even when the user can provide images which contain instances of the salient visual properties, content or configurations they would like to search for, it is very hard for the system to ascertain which aspects make a given image relevant and how similarity should be assessed. Many such systems therefore rely on extensive relevance feedback to guide the search towards desirable images, but this approach is not appropriate for most real-world retrieval scenarios. Many industrial applications of CBIR require ways of succinctly expressing abstract requirements which can not be encapsulated by any particular sample image.

- *Template, region selection, or sketch*: ([7], [23], [6]) Rather than providing whole images, the user can draw (sometimes literally) the system's attention to particular image aspects such as the spatial composition of desired content in terms of particular regions or a set of pre-defined templates. Clearly this process becomes cumbersome for complex queries and there are difficult user interface issues concerning how one might best represent abstract relations and invariants.

- *Feature range or predicate*: ([34], [33]) Here the user can set target ranges or thresholds for certain (typically low-level) attributes such as colour, shape, or texture features which may represent global image properties or features localised to certain image regions. While this clearly has merit for some types of queries, the approach requires a certain amount of user sophistication and patience.

- *Annotation or document context*: ([28], [46], [20]) Images rarely come with usable annotations for reasons such as cost, ambiguity, inconsistency, and human error. While progress has been made in applying text retrieval methods to annotations and other sources of image context such as captions, difficulties remain due to lack of availability, unreliability, and variability of such textual information.

- *Query language or concept*: ([8], [32], [41]) Efforts have been made to extend popular database query languages derived from SQL to cater for the intrinsic uncertainty involved in matching image features to assess relevance. However, such languages remain quite formal and rigid and are difficult to extend to higher-level concepts. Knowledge-based approaches utilising description logics or semantic networks have been proposed as a means of better representing semantic concepts but tend to entail somewhat cumbersome query interfaces.

Although these approaches have proven to be useful, both in isolation and when combined, in providing usable CBIR solutions for particular application domains and retrieval scenarios, much work remains to be done in providing query mechanisms that will scale and generalise to the applications envisaged for future mainstream content based access to multimedia. Indeed one criticism one can generally level at image retrieval systems is the extent to which they require the user to model the notions of content representation and similarity employed by the system rather than vice versa. One reason for the failure of CBIR to gain widespread adoption is due to the fact that mainstream users are quite unwilling to invest great effort into query composition [37] as many systems fail to perform in accordance with user expectations.

The language based query framework proposed in this paper aims to address these challenges. Query sentences are typically short (e.g. "people in centre") yet conceptually rich. This is because they need only represent those aspects of the target image(s) which the user is trying to retrieve and which distinguish such images from others in the dataset. The user is therefore not required to translate a description of an envisaged target image into the language but merely (and crucially) to express desired properties which are to hold for the retrieved images.

1.2 Language-based Querying

This paper argues that query languages constitute an important avenue for further work in developing CBIR query mechanisms. Powerful and easy-to-use textual document retrieval systems have become pervasive and constitute one of the major driving forces behind the internet. Given that so many people are familiar with the use of simple keyword strings and regular expressions to retrieve documents from vast online collections, it seems natural to extend language based querying to multimedia data. However, it is important to recognise [47] that the natural primitives of document retrieval, words and phrases, carry with them inherently more semantic information and characterise document content in a much more redundant and high level way than the pixels and simple features found in images. This is why text retrieval has been so successful despite the relative simplicity of using statistical measures to represent content *indicatively* rather than *substantively*. Image retrieval addresses a much more complex and ambiguous challenge, which is why we argue strongly for a query method based on a language that can represent both the *syntax* and *semantics* of image content at different conceptual levels. This paper will show that by basing this language on an ontology one can capture both concrete and abstract relationships between salient image properties such as objects in a much more powerful way than with the relatively weak co-occurrence based knowledge representation facilities of classical information retrieval. Since the language is used to express queries rather than describe image content, such relationships can be represented explicitly without prior commitments to a particular interpretation or having to incur the combinatorial explosion of an exhaustive annotation of all the relations that may hold in a given image. Instead, only those image aspects which are of value in determining relevance given a particular query are evaluated and evaluation may stop as soon as an image can be deemed irrelevant.

The comparatively small number of query languages designed for CBIR have largely failed to attain the standards necessary for general adoption. A major reason for this is the fact that most language or text based image retrieval systems rely on manual annotations, captions, document context, or pre-generated keywords, which leads to a loss of flexibility through the initial choice of annotation and indexing. Languages mainly concerned with deriving textual descriptions of image content [1] are inappropriate for general purpose retrieval since it is infeasible to generate exhaustive textual representations which contain all the information and levels of detail which might be required to process a given query in light of the user's retrieval need. While keyword indexing of images in terms of descriptors for semantic content remains highly desirable, semi- or fully automated annotation is currently based on image document context [42] or limited to low level descriptors. More ambitious "user-in-the-loop" annotation systems still require a substantial amount of manual effort to derive meaningful annotations [51]. Formal query languages such as extensions of SQL [38] are limited in their expressive power and extensibility and require a certain level of user experience and sophistication.

In order to address the challenges mentioned above while keeping user search overheads to a minimum, we have developed the *OQUEL* query description language. It provides an extensible language framework based on a formally specified grammar and an extensible vocabulary which are derived from a general ontology of image content in terms of categories, objects, attributes, and relations. Words in the language represent predicates on image features and target content at different semantic levels and serve as nouns, adjectives, and prepositions. Sentences are prescriptions of desired characteristics which are to hold for relevant retrieved images. They can represent spatial, object compositional, and more abstract relationships between terms and sub-sentences. The language is portable to other image content representation systems in that the lower level words and the evaluation functions which act on them can be changed or re-implemented with little or no impact on the conceptually higher language elements. It is also extensible since new terms can be defined both on the basis of existing constructs and based on new sources of image knowledge and metadata. This enables definition of customised ontologies of objects and abstract relations.

2 Ontological Language Framework

2.1 Role of Ontologies

By basing a retrieval language on an ontology, one can explicitly encode ontological commitments about the domain of interest in terms of categories, objects, attributes, and relations. Gruber [16] defines the term ontology in a knowledge sharing context as a "formal, explicit specification of a shared conceptualisation". Ontologies encode the relational structure of concepts which one can use to describe and reason about aspects of the world. Ontology is the theory of objects in terms of the criteria which allow one to distinguish between different types of objects and the relations, dependencies, and properties through which they may be described. For the present purposes ontologies are representations of image content at different semantic levels and queries expressing desired image characteristics.

Sentences in a language built by means of an ontology can be regarded as active representational constructs of information as knowledge and there have been similar approaches in the past applying knowledge-based techniques such as description logics ([18], [2], [3]) to CBIR. However, in many such cases the knowledge based relational constructs are simply translated into equivalent database query statements such as SQL [19] or a potentially expensive software agent methodology is employed for the retrieval process [10]. This mapping of ontological structures onto real-world evidence can be implemented in a variety of ways. Common approaches are heavily influenced by methods such as description logics, frame-based system, and Bayesian inference [13].

This paper argues that the role of a query language for CBIR should be primarily *prescriptive*, i.e. a sentence is regarded as a description of a user's retrieval requirements which cannot easily be mapped onto the description of image content available to the system. While the language presented here is designed from a general ontology which determines its lexical and syn-

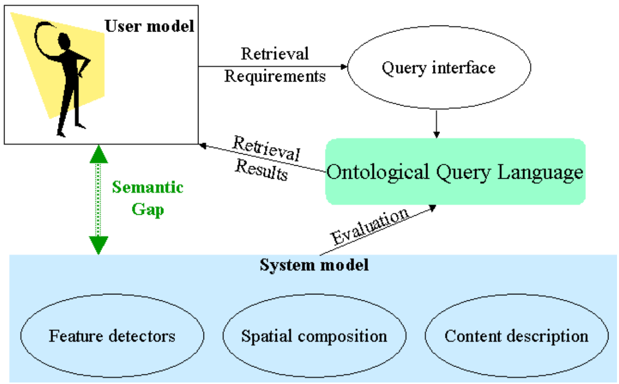


Figure 1: **Model of the retrieval process using an ontological query language to bridge the semantic gap between user and system notions of content and similarity.**

tactic elements to represent objects, attributes, and relations, this does not in itself constitute a commitment to a particular scheme for determining the semantic interpretation of any given query sentence. The evaluation of queries will depend on the makeup of the query itself, the indexing information available for each image, and the overall retrieval context. Evaluation therefore takes place within a particular ontological domain specified by the composition of the query and the available image evidence at the time it is processed. This approach is consistent with the view expressed in e.g. [41] that the *meaning* of an image is an emergent property which depends on both the query context and the image set over which the query is posed. Figure 1 shows how the ontological query language and the mechanisms for its interpretation can thus be regarded as acting as an intermediary between user and retrieval system in order to reduce the semantic gap.

2.2 Query Specific Image Interpretation

The important distinction between *query description* and *image description* languages is founded on the principle that while a given picture may well say more than a thousand words, a short query sentence expressed in a sufficiently powerful language can adequately describe those image properties which are relevant to a particular query. Information theoretic measures can then be applied to optimise a given query by identifying those of its elements which have high discriminative power to iteratively narrow down the search to a small number of candidate images. Hence it is the query itself which is taken as evidence for the relevance assessment measures appropriate to the user's retrieval requirement and "point of view". The syntax and semantics of the query sentence composed by the user thereby define the particular ontological domain in which the search for relevant images takes place. This is inherently a far more powerful way of relating image semantics to user requests than static image annotation which, even when carried out by skilled human annotators, will always fall far short of encapsulating those aspects and relationships which are of particular value in characterising an image in light of a new query.

The use of ontologies also offers the advantage of bridging between high-level concepts and low-level primitives in a way which allows extensions to the language to be defined on the basis of existing constructs without having to alter the representation of image data. Queries can thus span a range of conceptual granularity from concrete image features (regions, colour, shape, texture) and concrete relations (feature distance, spatial proximity, size) to abstract content descriptors (objects, scene descriptions) and abstract relations (similarity, class membership, inferred object and scene composition). The ability to automatically infer the presence of high-level concepts (e.g. a beach scene) on the basis of evidence (colour, region classification, composition) requires techniques such as Bayesian inference which plays an increasing role in semantic content derivation [50]. By expressing the causal relationships used to integrate multiple sources of evidence and content modalities in a dependency graph, such methods are also of great utility in quickly eliminating improbable configurations and thus narrowing down the search to a rapidly decreasing number of images which are potentially relevant to the query.

3 Language Design and Structure

This section introduces the *OQUEL* ontological query language with particular reference to its current implementation as a query description language for the ICON content based image retrieval system. For reasons of clarity, only a high level description of the language will be presented here. Section 4 will discuss implementation details pertaining to the content extraction and representation schemes used in the system and show how tokens in the language are mapped onto concrete image properties. Section 5 will show how query sentences are processed to assess image relevance.

3.1 Overview and Design Principles

The primary aim in designing OQUEL has been to provide both ordinary users and professional image archivists with an intuitive and highly versatile means of expressing their retrieval requirements through the use of familiar natural language words and a straightforward syntax. As mentioned above, many query languages have traditionally followed a path set out by database languages such as SQL which are characterised by a fairly sparse and restrictive grammatical framework aimed at facilitating concise and well-defined queries. The advantages of such an approach are many, e.g. ease of machine interpretation, availability of query optimisation techniques, scalability, theoretical analysis, etc.. However, their appropriateness and applicability to a domain of such intrinsic ambiguity and uncertainty as image retrieval remain doubtful. OQUEL was therefore designed to provide greater naturalness and flexibility through the use of a more complex grammar bearing a resemblance to natural language on a restricted domain.

3.2 Syntax and Semantics

In order to allow users to enter both simple keyword phrases and arbitrarily complex compound queries, the language grammar features constructs such as predicates, relations, conjunctions, and a specification syntax for image content. The latter includes adjectives for image region properties (i.e. shape, colour, and texture) and both relative and absolute object location. Desired image content can be denoted by nouns such as labels for automatically recognised visual categories of stuff ("grass", "cloth", "sky", etc.) and through the use of derived higher level terms for composite objects and scene description (e.g. "animals", "vegetation", "winter scene"). This includes the simple morphological distinction between singular and plural forms of certain terms, hence "people" will be evaluated differently from "person".

Tokens serving as adjectives denoting desired image properties are parameterised to enable values and ranges to be specified. The use of defaults, terms representing fuzzy value sets, and simple rules for operator precedence and associativity help to reduce the effective complexity of query sentences and limit the need for special syntax such as brackets to disambiguate grouping. Brackets can however optionally be used to define the scope of the logical operators (not, and, or, xor) and are required in rare cases to prevent the language from being context sensitive in the grammar theory sense.

While the inherent sophistication of the OQUEL language enables advanced users to specify extremely detailed queries if desired, much of this complexity is hidden by a versatile query parser. The parser was constructed with the aid of the SableCC lexer/parser generator tool from LALR(1) grammar rules and the WordNet [27] lexical database as further described in the next section. The vocabulary of the language is based on an annotated thesaurus of several hundred natural language words, phrases, and abbreviations (e.g. "!" for "not", "," for "and") which are recognised as tokens. Token recognition takes place in a lexical analysis step prior to syntax parsing to reduce the complexity of the grammar. This also makes it possible to provide more advanced word-sense disambiguation and analysis of phrasal structure while keeping the language efficiently LALR(1) parsable.

The following gives a somewhat simplified high level context free EBNF-style grammar G of the OQUEL language as currently implemented in the ICON system (capitals denote lexical categories, lower case strings are tokens or token sets).

$$
\begin{array}{rcl}
\multicolumn{3}{l}{G : \{} \\
S & \rightarrow & R \\
R & \rightarrow & modifier?\ (scenedescriptor \mid SB \mid BR) \\
 & & \mid not?\ R\ (CB\ R)? \\
BR & \rightarrow & SB\ binaryrelation\ SB \\
SB & \rightarrow & (CS \mid PS) +\ LS * \\
CS & \rightarrow & visualcategory \mid semanticcategory \mid \\
 & & not?\ CS\ (CB\ CS)? \\
LS & \rightarrow & location \mid not?\ LS\ (CB\ LS)? \\
PS & \rightarrow & shapedescriptor \mid colourdescriptor \mid \\
 & & sizedescriptor \mid not?\ PS\ (CB\ PS)? \\
CB & \rightarrow & and \mid or \mid xor; \\
\} & &
\end{array}
$$

The major syntactic categories are:

- *S*: start symbol of the sentence (text query)
- *R*: requirement (a query consists of one or more requirements which are evaluated separately, the probabilities of relevance then being combined according to the logical operators)
- *BR*: binary relation on SBs
- *SB*: specification block consisting of at least one CS or PS and 0 or more LS
- *CS*: image content specifier
- *LS*: location specifier for regions meeting the CS/PS
- *PS*: region property specifier (visual properties of regions such as colour, shape, texture, and size)
- *CB*: binary (fuzzy) logical connective (conjunction, disjunction, and exclusive-OR)

Tokens (terminals) belong to the following sets:

- *modifier*: Quantifiers such as "a lot of", "none", "as much as possible".
- *scene descriptor*: Categories of image content characterising an entire image, e.g. countryside, city, indoors.
- *binaryrelation*: Relationships which are to hold between clusters of target content denoted by specification blocks. The current implementation includes spatial relationships such as "larger than", "close to", "similar size as", "above", etc. and some more abstract relations such as "similar content".
- *visualcategory*: Categories of stuff, e.g. water, skin, cloud.
- *semanticcategory*: Higher semantic categories such as people, vehicles, animals.
- *location*: Desired location of image content matching the content or shape specification, e.g. "background", "lower half", "top right corner".
- *shapedescriptor*: Region shape properties, for example "straight line", "blob shaped".
- *colourdescriptor*: Region colour specified either numerically or through the use of adjectives and nouns, e.g. "bright red", "dark green", "vivid colours".
- *sizedescriptor*: Desired size of regions matching the other criteria in a requirement, e.g. "at least 10%" (of image area), "largest region".

The precise semantics of these constructs are dependent upon the way in which the query language is implemented, the parsing algorithm, and the user query itself, as will be described in the following sections.

3.3 Vocabulary

As shown in the previous section, OQUEL features a generic base vocabulary built on extracted image features and intermediate level content labels which can be assigned to segmented image regions on the basis of such features. Some terminal symbols of the language therefore correspond directly to previously extracted image descriptors. This base vocabulary has been extended and remains extensible by derived terms denoting higher level objects and concepts which can be inferred at query time. While the current OQUEL implementation is geared towards general purpose image retrieval from photographic image collections, task specific vocabulary extensions can also be envisaged.

In order to provide a rich thesaurus of synonyms and also capture some more complex relations and semantic hierarchies of words and word senses, we have made use of the lexical information present in the WordNet [27] electronic dictionary. This contains a large vocabulary which has been systematically annotated with word sense information and relationships such as synonyms, antonyms, hyper- and hyponyms, meronyms, etc.. Currently we have used some of this information to define a thesaurus of about 400 words relating to the extracted image features and semantic descriptors mentioned above.

Work has begun on improving the flexibility of the OQUEL retrieval language by adding a pre-processing stage to the current query parser. This will use additional semantic associations and word relationships encoded in the WordNet database to provide much greater expressive power and ease syntactical constraints. Such a move may require a more flexible natural-language oriented parsing strategy to cope with the additional difficulty of word-sense and query structure disambiguation but will also pave the way for future work on using the language as a powerful representational device for content-based knowledge extraction.

3.4 Example sentences

The following are examples of valid OQUEL queries as used in conjunction with ICON:

> some sky which is close to buildings in upper corner
>
> some water in the bottom half which is surrounded by trees and grass, size at least 10%
>
> [indoors] & [people]
>
> some green or vividly coloured vegetation in the centre which is of similar size as clouds or blue sky at the top
>
> [artificial stuff, vivid colours and straight lines] and tarmac

4 Implementation of the OQUEL Language

4.1 Content Extraction and Representation

ICON (Image Content Organisation and Navigation, [21]) combines a cross-platform Java user interface with image processing and content analysis functionality to facilitate automated organisation of and retrieval from large heterogeneous image sets based on both meta data and visual content.

The backend image processing components extract various types of content descriptors and meta data from images (see [49]). The following are currently used in conjunction with OQUEL queries:

Image segmentation: Images are segmented into non-overlapping regions and sets of properties for size, colour, shape, and texture are computed for each region [43, 44]. Initially full three colour edge detection is performed using the weighted total change dT

$$dT = dI_i^2 + dI_j^2 + 3.0dC \tag{1}$$

where the total change in intensity dI_i is given by the colour derivatives in RGB space

$$dI_i = dR_i + dG_i + dB_i \tag{2}$$

and the magnitude of change in colour is represented by

$$\begin{aligned} dC &= \sqrt{(}(dB_i - dG_i)^2 + (dR_i - dB_i)^2 + (dG_i - dR_i)^2 \\ &+ (dB_j - dG_j)^2 + (dR_j - dB_j)^2 + (dG_j - dR_j)^2) \end{aligned} \tag{3}$$

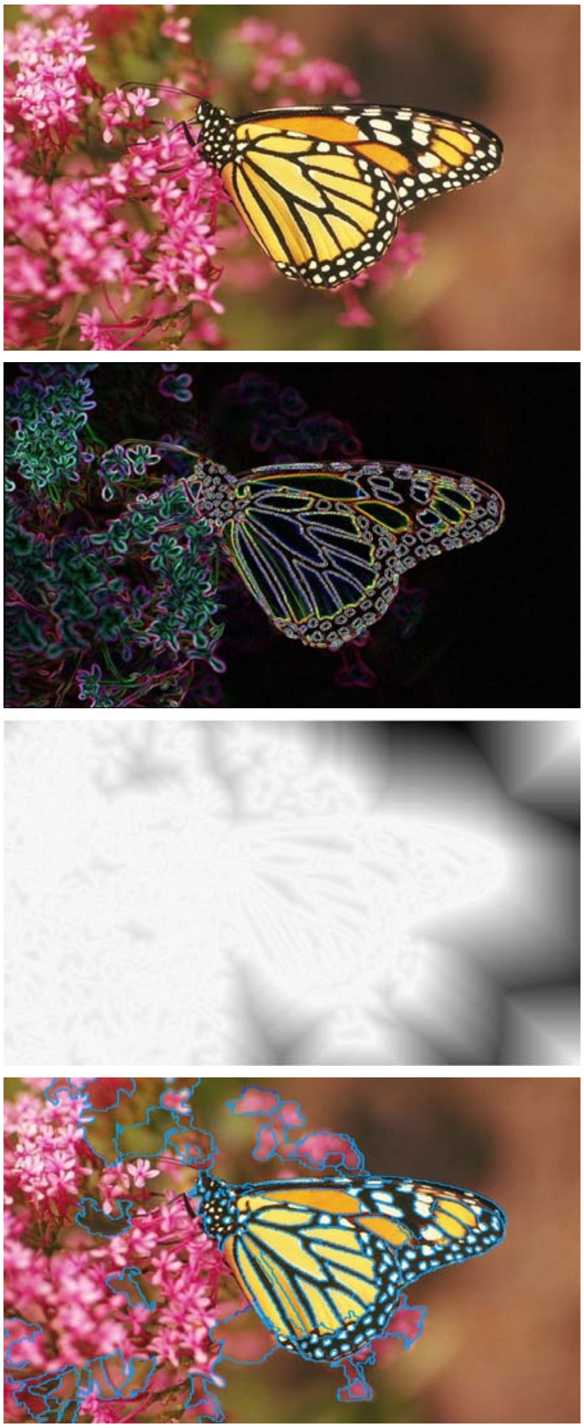


Figure 2: *From top to bottom: Example colour image from the Corel image library. Full three colour derivative of the image. Voronoi image computed from the edges found by the three colour edge detector (the darker a pixel, the further it is from an edge). Final region segmentation (boundaries indicated in blue).*

The choice of 3 as the weighting factor in favour of colour change over brightness change is empirical but has been found to be effective across a very broad range of photographic images and artwork. Local orientation (for use in the non-maximum suppression step of the edge detection process) is defined to be in the direction of the maximum colour gradient. dT is then the edge strength fed to the non-max suppression and hysteresis edge-following steps which follow the popular method due to *Canny* .

Voronoi seed points for region growing are generated from the peaks in the distance transform of the edge image, and regions are then grown agglomeratively from seed points with gates on colour difference with respect to the boundary colour and mean colour across the region. Unassigned pixels at the boundaries of a growing region are incorporated into a region if the difference in colour between it and pixels in the local neighbourhood of the region is less than one threshold and the difference in colour between the candidate and the mean colour of the region is less than a second larger threshold.

A texture model based on discrete ridge features is also used to describe regions in terms of texture feature orientation and density. Ridge pixels are those for which the magnitude of the second derivative operator applied to a grey-scale version of the original image exceeds a threshold. The network of ridges is then broken up into compact 30 pixel feature groups and the orientation of each feature is computed from the second moment about its center of mass. Features are clustered using Euclidean distance in RGB space and the resulting clusters are then employed to unify regions which share significant portions of the same feature cluster. The internal brightness structure of "smooth" (largely untextured) regions in terms of their isobrightness contours and intensity gradients is used to derive a parameterisation of brightness variation which allows shading phenomena such as bowls, ridges, folds, and slopes to be identified. A histogram representation of colour covariance and shape features is computed for regions above a certain size threshold.

The segmentation scheme then returns a region map together with internal region description parameters comprising colour, colour covariance, shape, texture, location and adjacency relations. Segmentation does not rely on large banks of filters to estimate local image properties and hence is fast (typically a few seconds for high resolution digital photographa) and does not suffer from the problem of the boundary between two regions appearing as a region itself. The region growing technique effectively surmounts the problem of broken edged topology and the texture feature based region unification step ensures that textured regions are not fractured. Figure 2 shows a sample image from the Corel picture library and the results of segmentation. The number of segmented regions depends on image size and visual content, but has the desirable property that most of the image area is commonly contained within a few dozen regions which closely correspond to the salient features of the picture at the conceptual granularity of the semantic categories used here.

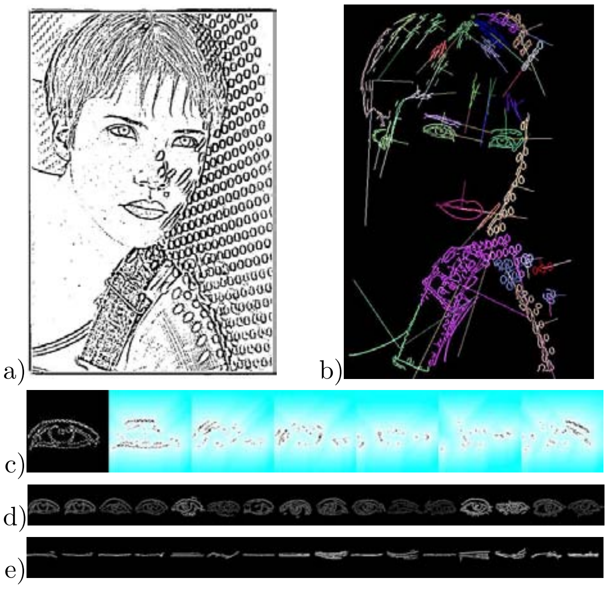


Figure 3: *Eye detection process: a) Binarised image. b) Hausdorff clustered regions after filtering. c) Normalised feature cluster of the left eye (left) and distance transforms for 6 feature orientations (blue areas are further from feature points). d) Examples of left eyes correctly classified using nearest neighbours. e) Examples of nearest neighbour clusters for non-eyes.*

Stuff classification: Region descriptors computed from the segmentation algorithm are fed into artificial neural network classifiers which have been trained to label regions with class membership probabilities for a set of 12 semantically meaningful visual categories of "stuff" ("Brick", "Blue sky", "Cloth", "Cloudy sky", "Grass", "Internal walls", "Skin", "Snow", "Tarmac", "Trees", "Water", and "Wood"). The classifiers are MLP (multi layer perceptron) and RBF (radial basis function) networks whose topology was optimised to yield best generalisation performance for each particular visual category using separate training, testing and validation sets from a large (over 40000 exemplars) corpus of manually labelled image regions. The MLP networks typically consist of three hidden layers with progressively smaller numbers of neurons (up to 250) in each layer.

Automatic labelling of segmented image regions with semantic visual categories [49] such as grass or water which mirror aspects of human perception allows the implementation of intuitive and versatile query composition methods while greatly reducing the search space. The current set of categories was chosen to facilitate robust classification of general photographic images. These categories are by no means exhaustive but represent a first step towards identifying fairly low level semantic properties of image regions which can be used to ground higher level concepts and content prescriptions. Various psychophysical studies [9, 29] have shown that semantic descriptors such as these serve as useful cues for determining image content by humans and CBIR systems. An attempt was made to include categories which allow one to distinguish between indoor and outdoor scenes. Experiments on both the Corel photo set and a large body of amateur digital photographs have given classification success rates of between 82% and 96% for the largest image regions which jointly cover over 80% of image area.

Colour descriptors: Nearest-neighbour colour classifiers were built from the region colour representation. These use the Earth-mover distance measure applied to Euclidean distances in RGB space to compare region colour profiles with cluster templates learned from a training set. In a manner similar to related approaches such as [34, 29], colour classifiers were constructed for each of twelve “basic” colours (“black”, “blue”, “cyan”, “grey”, “green”, “magenta”, “orange”, “pink”, “red”, “white”, “yellow”, “brown”). Each region is associated with the colour labels which best describe it.

Face detection: Face detection relies on identifying elliptical regions (or clusters of regions) classified as human skin. A binarisation transform is then performed on a smoothed version of the image. Candidate regions are clustered based on a Hausdorff distance measure [39] and resulting clusters are filtered by size and overall shape and normalised for orientation and scale. From this a spatially indexed oriented shape model is derived by means of a distance transform of 6 different orientations of edge-like components from the clusters via pairwise geometric histogram binning [12]. A nearest-neighbour shape classifier was trained to recognise eyes. See figure 3 for an illustration of the approach.

Adjacent image regions classified as human skin in which eye candidates have been identified are then labelled as containing (or being part of) one or more human faces subject to the scale factor implied by the separation of the eyes. This detection scheme shows robustness across a large range of scales, orientations, and lighting conditions but suffers from false positives. Recently an integrated face detector based on a two level classifier of polynomial kernel SVMs (Support Vector Machines) has been implemented. For reasons of efficiency, this detector is applied only to face candidates detected by the previously described method in order to greatly reduce the false positive rate while retaining high accuracy (about 72% correct detections and 0.08 false positives per image for the diverse image collection employed in section 6).

Region mask: Canonical representation of the segmented image giving the absolute location of each region by mapping pixel locations onto region identifiers. The mask stores an index value into the array of regions in order to indicate which region each pixel belongs to. For space efficiency this is stored in a run length encoded representation.

Region graph: Graph of the relative spatial relationships of the regions (distance, adjacency, joint boundary, and containment). Distance is defined in terms of the Euclidean distance between centres of gravity, adjacency is a binary property denoting that regions share a common boundary segment, and the joint boundary property gives the relative proportion of region boundary shared by adjacent regions. A region A is said to be contained within region B if A shares 100% of its boundary with B. Together with the simple parameterisation of region shape computed by the segmentation method, this provides an efficient (if non-exact) representation of the geometric relationships between image regions.

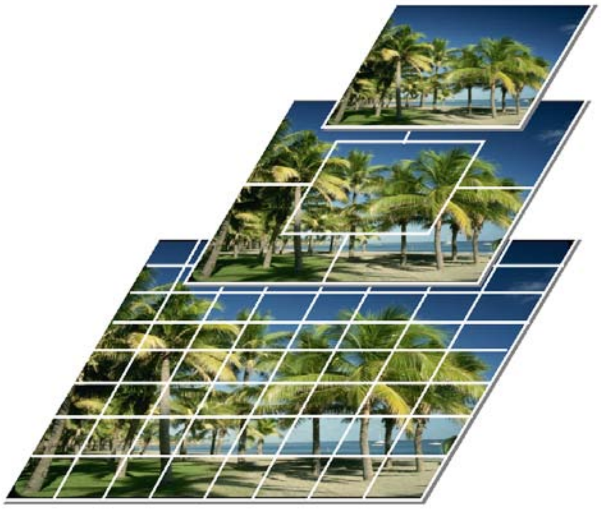


Figure 4: **Three level grid pyramid which subdivides an image into different numbers of fixed grids (1, 5, 64) at each level.**

Grid pyramid: The proportion of image content which has been positively classified with each particular label (visual category, colour, and presence of faces) at different levels of an image pyramid (whole image, image fifths, 8x8 chess grid, see figure 4). For each grid element we therefore have a vector of percentages for the 12 stuff categories, the 12 colour labels, and the percentage of content deemed to be part of a human face. Grid regions are generally of the same area and rectangular shape, except in the case of the image fifths where the central rectangular fifth occupies 25% of image area and is often given a higher weighting for scene characterisation to reflect the fact that this region is likely to constitute the most salient part of the image. Through the relationship graph representation of regions we can make the matching of clusters of regions invariant with respect to displacement and rotation using standard matching algorithms [36]. The grid pyramid and region mask representations allow an efficient comparison of absolute position and size.

This may be regarded as an intermediate level representation which does not preclude additional stages of visual inference and composite object recognition in light of query specific saliency measures and the integration of contextual information. Such intermediate level semantic descriptors for image content have been used by several CBIR systems in recent years ([26], [5], [14], [22]).

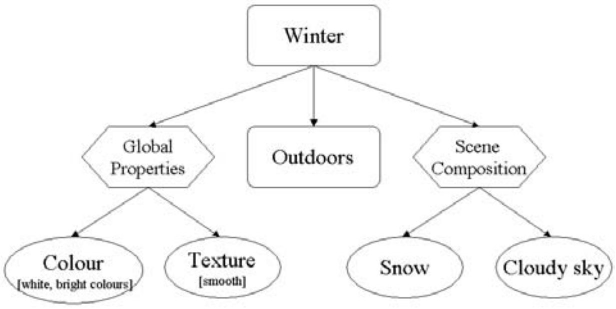


Figure 5: **Simplified Bayesian network for the scene descriptor "winter".**

4.2 Grounding the Vocabulary

An important aspect of OQUEL language implementation concerns the way in which sentences in the languages are *grounded* in the image domain. Here we discuss those elements of the token set which might be regarded as being statically grounded, i.e. there exists a straightforward mapping from OQUEL words to extracted image properties as described in section 4.1. Other terminals (modifiers, scene descriptors, binary relations, and semantic categories) and syntactic constructs are evaluated by the query parses as will be discussed in section 5.

visualcategory: The 12 categories of stuff which have been assigned to segmented image regions by the neural net classifiers. Assignment of category labels to image region is based on a threshold applied to the classifier output.

location: Location specifiers which are simply mapped onto the grid pyramid representation. For example, when searching for "grass" in the "bottom left" part of an image, only content in the lower left image fifth will be considered.

shapedescriptor: The current terms are "straight line", "vertical", "horizontal", "stripe", "right angle", "top edge", "left edge", "right edge", "bottom edge", "polygonal", and "blobs". They are defined as predicates over region properties and aspects of the region graph representation derived from the image segmentation. For example, a region is deemed to be a straight line if its shape is well approximated by a thin rectangle, "right edge" corresponds to a shape appearing along the right edge of the image, and "blobs" are regions with highly amorphous shape without straight line segments.

colourdescriptor: Region colour specified either numerically in the RGB or HSV colour space or through colour labels assigned by the nearest-neighbour classifiers. By assessing the overall brightness and contrast properties of a region using fixed thresholds, colours identified by each classifier can be further described by a set of three "colour modifiers" ("bright", "dark", "faded").

sizedescriptor: The size of image content matching other aspects of a query is assessed by adding the areas of the corresponding regions. Size may be defined as a percentage value of image area ("at least x%", "at most x%", "between x% and y%") or relative to other image parts (e.g. "largest", "smallest", "bigger than").

4.3 System Integration

A general query methodology for content based image and multimedia retrieval must take into account the differences in potential application domains and system environments. Great care was therefore taken in the design of the OQUEL language to make it possible to integrate it with existing database infrastructure and content analysis facilities. This *portability* was achieved by a component-based software development approach which allows individual matching modules to be re-implemented to cater for alternative content representation schemes without affecting the overall semantics of the query language. This facility also makes it possible to evaluate a particular query differently depending on the current retrieval context.

The implementation of OQUEL also remains *extensible*. New terms can be represented on the basis of existing constructs as macro definitions. Simple lexical extensions are handled by a tokeniser and do not require any modifications to the query parser. Novel concepts can also be introduced by writing an appropriate software module (a Java class extending an interface or derived by inheritance) and plugging it into the existing language model. While an extension of the language syntax requires recompilation of the grammar specification, individual components of the language are largely independent and may be re-specified without affecting other parts. Furthermore, translation modules can be defined to optimise query evaluation or transform part of the query into an alternative format (e.g. a sequence of pre-processed SQL statements).

As will be discussed in the next section, the query text parser was designed to hide grammatical complexity ("what you don't know can't hurt you") and provide a natural language like tool for query composition. There is also a forms-based interface which offers the look and feel of graphical database interfaces and explicitly exposes available language features while being slightly restricted in the type of queries it can handle. Lastly there is a graphical tool which allows users to inspect or modify a simplified abstract syntax tree (AST) representation of a query.

5 Retrieval Process

This section discusses the OQUEL retrieval process as implemented in the ICON system. In the first stage, the syntax tree derived from the query is parsed top down and the leaf nodes are evaluated in light of their predecessors and siblings. Information then propagates back up the tree until we arrive at a single probability of relevance for the entire image. At the lowest level, tokens map directly or very simply onto the content descriptors. Higher level terms are either expanded into sentence representations or evaluated using Bayesian graphs. For example, when looking for people in an image the system will analyse the presence and spatial composition of appropriate clusters of relevant stuff (cloth, skin, hair) and relate this to the output of face and eye spotters. This evidence is then combined probabilistically to yield a likelihood estimate of whether people are present in the image. Figure 5 shows a simplified Bayesian network for the scene descriptor "winter". Arrows denote conditional dependencies and terminal nodes correspond to sources of evidence or, in the case of the term "outdoors", other Bayesian nets.

5.1 Query-time Object and Scene Recognition for Retrieval

In this paper we have argued that in order to come closer to capturing the semantic "essence" of an image, tasks such as feature grouping and object identification need to be approached in an adaptive goal oriented manner. This takes into account that criteria for what constitutes non-accidental and perceptually significant visual properties necessarily depend on the objectives and prior knowledge of the observer, as recognised in [4]. Going back to the lessons learned from text retrieval stated in section 1.2, for most content retrieval tasks it is perfectly adequate to approach the problem of retrieving images containing particular objects or characterisable by particular scene descriptors in an *indicative* fashion rather than a full *analytic* one. As long as the structure of the inference methods adequately accounts for the non-accidental properties that characterise an object or scene, relevance can be assessed by a combination of individually weak sources of evidence. These can be ranked in a hierarchy and further divided into those which are *necessary* for the object to be deemed present and those which are merely *contingent*. Such a ranking makes it possible to quickly eliminate highly improbable images and narrow down the search window.

Relevant images are those where we can find sufficient support for the candidate hypotheses derived from the query. Given enough redundancy and a manageable false positive rate, this will be resilient to failure of individual detection modules. For example, a query asking for images containing people does not require us to solve the full object recognition challenge of correctly identifying the location, gender, size, etc. of all people depicted in all images in the collection. As long as we maintain a notion of uncertainty, borderline false detections will simply result in lowly ranked retrieved images. Top query results will correspond to those image where our confidence of having found evidence for the presence of people is high relative to the other images, subject to the inevitable thresholding and identification of necessary features.

5.2 Query Parsing and Representation

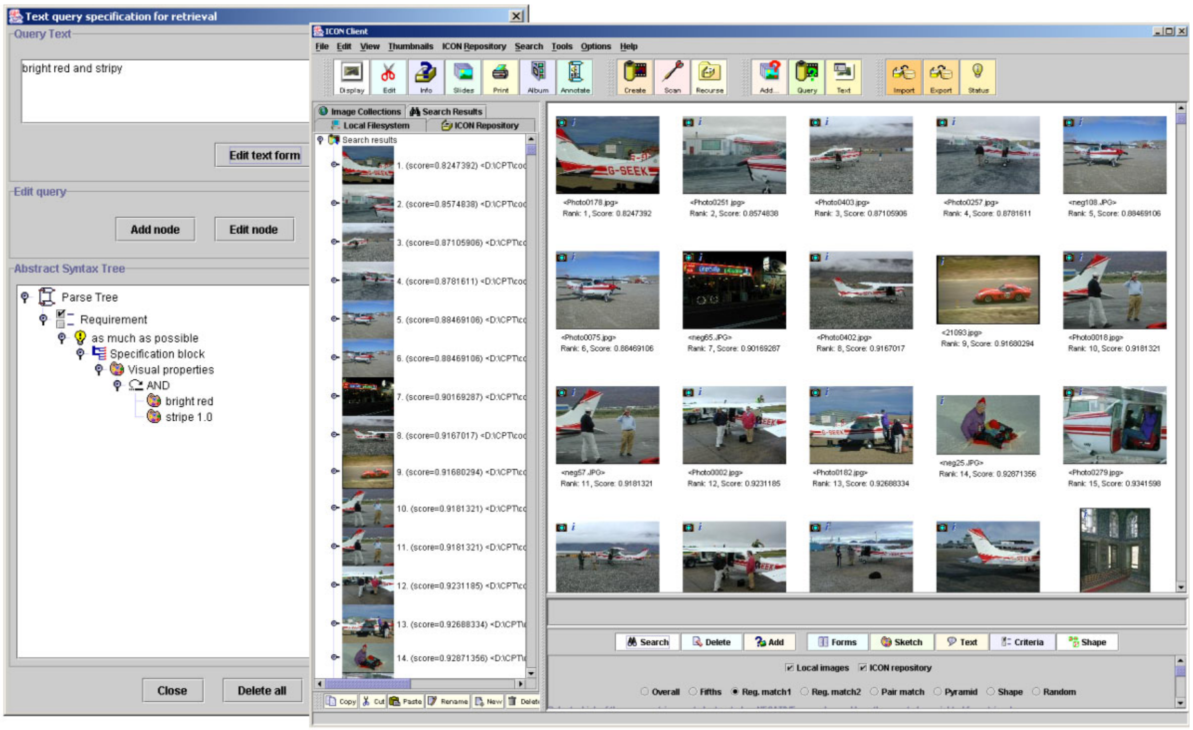
OQUEL queries are parsed to yield a canonical abstract syntax tree (AST) representation of their syntactic structure. Figures 6, 7, 8, and 9 show sample queries and their ASTs. The structure of the syntax trees follows that of the grammar, i.e. the root node is the start symbol whose children represent particular requirements over image features and content. The leaf nodes of the tree correspond to the terminal symbols representing particular requirements such as shapedescriptors and visual categories. Intermediate nodes are syntactic categories instantiated with the relevant token (i.e. "AND", "which is larger than") which represent the relationships that are to be applied when evaluating the query.

5.3 Query Evaluation and Retrieval

Images are retrieved by evaluating the AST to compute a probability of relevance for each image. Due to the inherent uncertainty and complexity of the task, evaluation is performed in a manner which limits the requirement for runtime inference by quickly ruling out irrelevant images given the query. Query sentences consist of requirements which yield matching probabilities that are further modified and combined according to the top level syntax. Relations are evaluated by considering the image evidence returned by assessing their constituent specification blocks. These attempt to find a set of candidate image content (evidence) labelled with probabilities according to the location, content, and property specifications which occur in the syntax tree. A closure consisting of a pointer to the identified content (e.g. a region identifier or grid coordinate) together with the probability of relevance is passed as a message to higher levels in the tree for evaluation and fusion.

The overall approach therefore relies on passing messages (image structures labelled with probabilities of relevance), assigning weights to these messages according to higher level structural nodes (modifiers and relations), and integrating these at the topmost levels (specification blocks) in order to compute a belief state for the relevance of the evidence extracted from the given image for the given query. There are many approaches to using probabilities to quantify and combine uncertainties and beliefs in this way [35]. The approach adopted here is related to that of [25] in that it applies notions of weighting derived from the Dempster-Shafer theory of evidence to construct an information retrieval model which captures structure, significance, uncertainty, and partiality in the evaluation process.

At the leaf nodes of the AST, derived terms such as object labels ("people") and scene descriptions ("indoors") are either expanded into equivalent OQUEL sentence structures or evaluated by Bayesian networks integrating image content descriptors with additional sources of evidence (e.g. a face detector). Bayesian networks tend to be context dependent in their applicability and may therefore give rise to brittle performance when applied to very general content labelling tasks. In the absence of additional information in the query sentence itself, it was therefore found useful to evaluate mutually exclusive scene descriptors for ad-


Text query specification for retrieval
ICON Client
File Edit View Thumbnails ICON Repository Search Tools Options Help
Display Edit Info Slides Print Album Annotate Create Scan Recurse Add... Query Text Import Export Status
Image Collections
Search Results
Local Filesystem
ICON Repository
Copy Cut Paste Rename New Delete
Search Delete Add Forms Sketch Text Criteria Shape
Edit text form
Add node
Edit node
Close
Delete all

Query Text

bright red and stripy

Edit query

Abstract Syntax Tree

- Parse Tree
  - Requirement
    - as much as possible
      - Specification block
        - Visual properties
          - AND
            - bright red
            - stripe 1.0

Search results

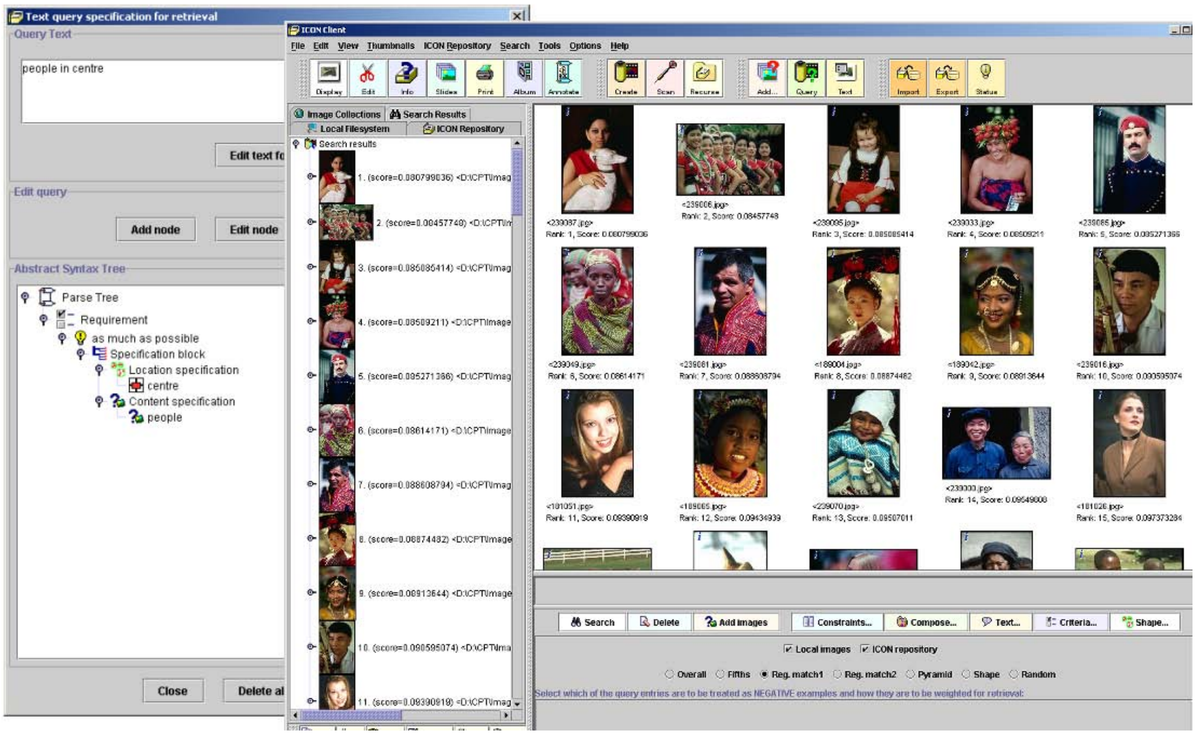
1. (score=0.8247392) <D:\CPT\coo
2. (score=0.8574838) <D:\CPT\coo
3. (score=0.87105906) <D:\CPT\co
4. (score=0.8781611) <D:\CPT\coo
5. (score=0.88469106) <D:\CPT\co
6. (score=0.88469106) <D:\CPT\co
7. (score=0.90169287) <D:\CPT\co
8. (score=0.9167017) <D:\CPT\coo
9. (score=0.91680294) <D:\CPT\co
10. (score=0.9181321) <D:\CPT\co
11. (score=0.9181321) <D:\CPT\co
12. (score=0.9231185) <D:\CPT\co
13. (score=0.92688334) <D:\CPT\c
14. (score=0.92871356) <D:\CPT\c

<Photo0178.jpg> Rank: 1, Score: 0.8247392
<Photo0251.jpg> Rank: 2, Score: 0.8574838
<Photo0403.jpg> Rank: 3, Score: 0.87105906
<Photo0257.jpg> Rank: 4, Score: 0.8781611
<neg108.JPG> Rank: 5, Score: 0.88469106
<Photo0075.jpg> Rank: 6, Score: 0.88469106
<neg65.JPG> Rank: 7, Score: 0.90169287
<Photo0402.jpg> Rank: 8, Score: 0.9167017
<21093.jpg> Rank: 9, Score: 0.91680294
<Photo0018.jpg> Rank: 10, Score: 0.9181321
<neg57.JPG> Rank: 11, Score: 0.9181321
<Photo0002.jpg> Rank: 12, Score: 0.9231185
<Photo0182.jpg> Rank: 13, Score: 0.92688334
<neg25.JPG> Rank: 14, Score: 0.92871356
<Photo0279.jpg> Rank: 15, Score: 0.9341598

☑ Local images ☑ ICON repository

○ Overall ○ Fifths ● Reg. match1 ○ Reg. match2 ○ Pair match ○ Pyramid ○ Shape ○ Random

Figure 6: **Search results for OQUEL query A "bright red and stripy".**


Text query specification for retrieval
ICON Client
File Edit View Thumbnails ICON Repository Search Tools Options Help
Display Edit Info Slides Print Album Annotate Create Scan Recurse Add... Query Text Import Export Status
Image Collections
Search Results
Local Filesystem
ICON Repository
Search Delete Add images Constraints... Compose... Text... Criteria... Shape...
Edit text fo
Add node
Edit node
Close
Delete al

Query Text

people in centre

Edit query

Abstract Syntax Tree

- Parse Tree
  - Requirement
    - as much as possible
      - Specification block
        - Location specification
          - centre
        - Content specification
          - people

Search results

1. (score=0.080799036) <D:\CPT\imag
2. (score=0.08457748) <D:\CPT\im
3. (score=0.085085414) <D:\CPT\imag
4. (score=0.08509211) <D:\CPT\image
5. (score=0.085271366) <D:\CPT\imag
6. (score=0.08614171) <D:\CPT\image
7. (score=0.088608794) <D:\CPT\imag
8. (score=0.08874482) <D:\CPT\image
9. (score=0.08913644) <D:\CPT\image
10. (score=0.090595074) <D:\CPT\ima
11. (score=0.09390919) <D:\CPT\imag

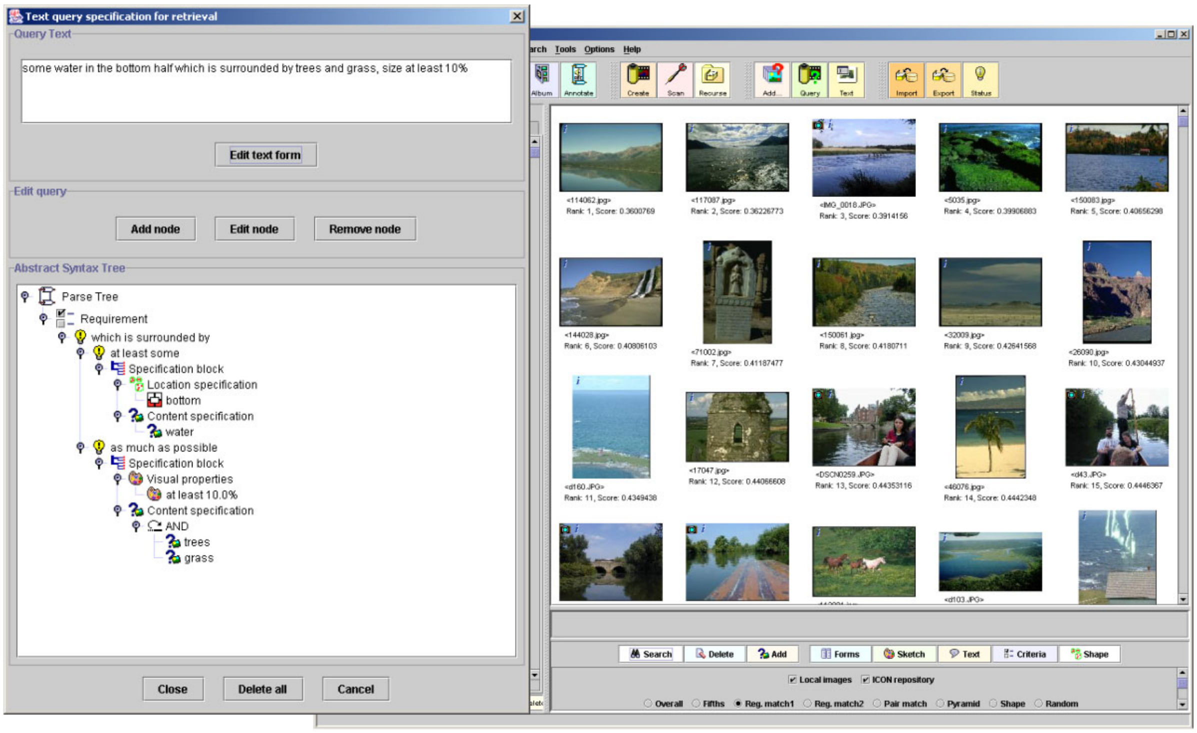
<239087.jpg> Rank: 1, Score: 0.080799036
<239006.jpg> Rank: 2, Score: 0.08457748
<239095.jpg> Rank: 3, Score: 0.085085414
<239033.jpg> Rank: 4, Score: 0.08509211
<239085.jpg> Rank: 5, Score: 0.085271366
<239049.jpg> Rank: 6, Score: 0.08614171
<239061.jpg> Rank: 7, Score: 0.088608794
<189004.jpg> Rank: 8, Score: 0.08874482
<189042.jpg> Rank: 9, Score: 0.08913644
<239016.jpg> Rank: 10, Score: 0.090595074
<181051.jpg> Rank: 11, Score: 0.09390919
<189065.jpg> Rank: 12, Score: 0.09434939
<239070.jpg> Rank: 13, Score: 0.09507011
<239000.jpg> Rank: 14, Score: 0.09549000
<181026.jpg> Rank: 15, Score: 0.097373284

☑ Local images ☑ ICON repository

○ Overall ○ Fifths ● Reg. match1 ○ Reg. match2 ○ Pyramid ○ Shape ○ Random

Select which of the query entries are to be treated as NEGATIVE examples and how they are to be weighted for retrieval.

Figure 7: **Search results for OQUEL query B "people in centre".**



Query Text: some water in the bottom half which is surrounded by trees and grass, size at least 10%

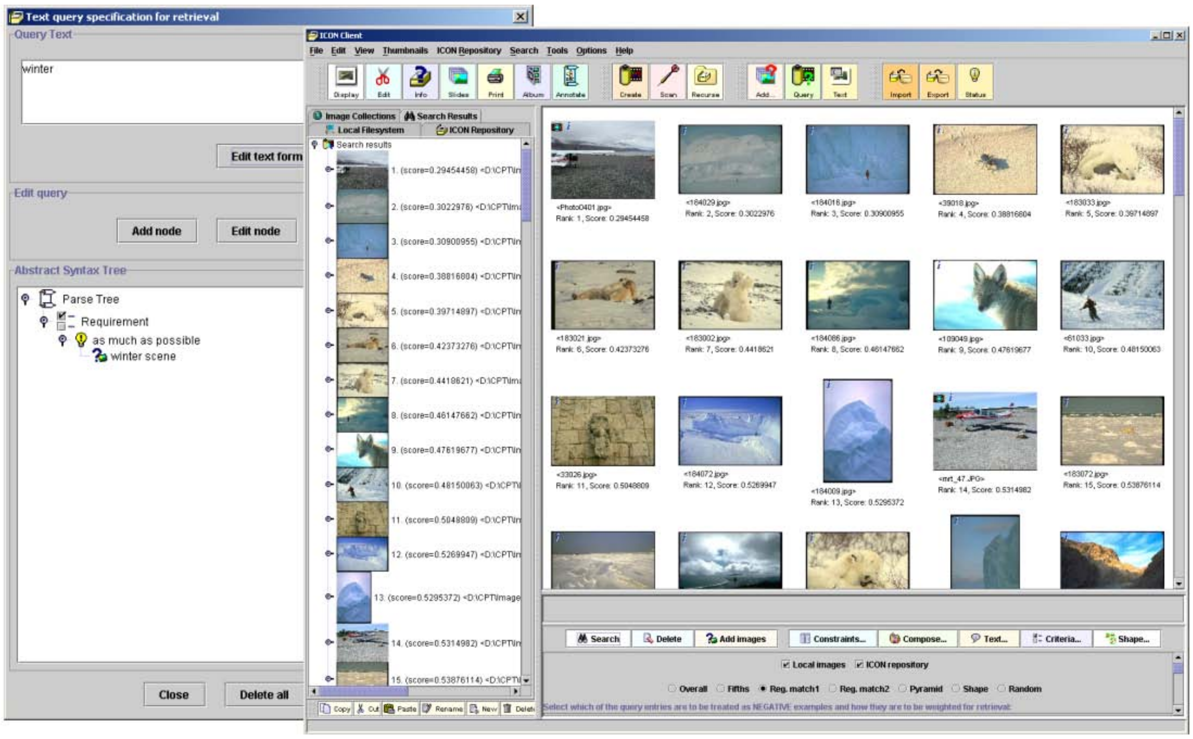
- Parse Tree
  - Requirement
    - which is surrounded by
      - at least some
        - Specification block
          - Location specification
            - bottom
          - Content specification
            - water
      - as much as possible
        - Specification block
          - Visual properties
            - at least 10.0%
          - Content specification
            - AND
              - trees
              - grass

<114062.jpg> Rank: 1, Score: 0.3600769
<117087.jpg> Rank: 2, Score: 0.36226773
<IMG_0018.JPG> Rank: 3, Score: 0.3914156
<5035.jpg> Rank: 4, Score: 0.39906883
<150083.jpg> Rank: 5, Score: 0.40656298
<144028.jpg> Rank: 6, Score: 0.40806103
<71002.jpg> Rank: 7, Score: 0.41187477
<150061.jpg> Rank: 8, Score: 0.4180711
<32009.jpg> Rank: 9, Score: 0.42641568
<26090.jpg> Rank: 10, Score: 0.43044937
<d160.JPG> Rank: 11, Score: 0.4349438
<17047.jpg> Rank: 12, Score: 0.44066608
<DSCN0259.JPG> Rank: 13, Score: 0.44353116
<46076.jpg> Rank: 14, Score: 0.4442348
<d43.JPG> Rank: 15, Score: 0.4446367
<d103.JPG>

☑ Local images ☑ ICON repository

○ Overall ○ Fifths ● Reg. match1 ○ Reg. match2 ○ Pair match ○ Pyramid ○ Shape ○ Random

Figure 8: **Search results for OQUEL query C "some water in the bottom half which is surrounded by trees and grass, size at least 10%".**



Query Text: winter

- Parse Tree
  - Requirement
    - as much as possible
      - winter scene

Search results
1. (score=0.29454458) <D:\CPT\Im
2. (score=0.3022976) <D:\CPT\Ima
3. (score=0.30900955) <D:\CPT\Im
4. (score=0.38816804) <D:\CPT\Im
5. (score=0.39714897) <D:\CPT\Im
6. (score=0.42373276) <D:\CPT\Im
7. (score=0.4418621) <D:\CPT\Ima
8. (score=0.46147662) <D:\CPT\Im
9. (score=0.47619677) <D:\CPT\Im
10. (score=0.48150063) <D:\CPT\
11. (score=0.5048809) <D:\CPT\Im
12. (score=0.5269947) <D:\CPT\Im
13. (score=0.5295372) <D:\CPT\Image
14. (score=0.5314982) <D:\CPT\Im
15. (score=0.53876114) <D:\CPT\

<Photo0401.jpg> Rank: 1, Score: 0.29454458
<184029.jpg> Rank: 2, Score: 0.3022976
<184016.jpg> Rank: 3, Score: 0.30900955
<39018.jpg> Rank: 4, Score: 0.38816804
<183033.jpg> Rank: 5, Score: 0.39714897
<183021.jpg> Rank: 6, Score: 0.42373276
<183002.jpg> Rank: 7, Score: 0.4418621
<184086.jpg> Rank: 8, Score: 0.46147662
<109049.jpg> Rank: 9, Score: 0.47619677
<61033.jpg> Rank: 10, Score: 0.48150063
<33026.jpg> Rank: 11, Score: 0.5048809
<184072.jpg> Rank: 12, Score: 0.5269947
<184009.jpg> Rank: 13, Score: 0.5295372
<mrt_47.JPG> Rank: 14, Score: 0.5314982
<183072.jpg> Rank: 15, Score: 0.53876114

☑ Local images ☑ ICON repository

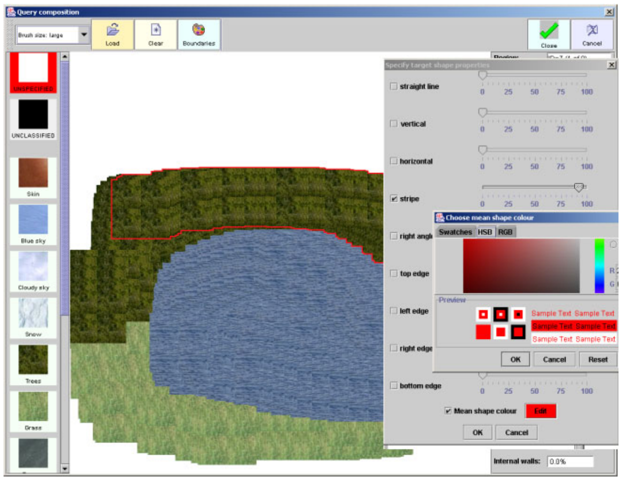
○ Overall ○ Fifths ● Reg. match1 ○ Reg. match2 ○ Pyramid ○ Shape ○ Random

Select which of the query entries are to be treated as NEGATIVE examples and how they are to be weighted for retrieval:

Figure 9: **Search results for OQUEL query D "winter".**

ditional disambiguation. For example, the concepts "winter" and "summer" are not merely negations of one another but correspond to Bayesian nets evaluating different sources of evidence. If both were to assign high probabilities to a particular image then the labelling is considered ambiguous and consequently assigned a lower relevance weight.

The logical connectives are evaluated using thresholding and fuzzy logic (i.e. "p1 and p2" corresponds to "if (min(p1,p2)<=threshold) 0 else min(p1,p2)"). A similar approach is taken in evaluating predicates for low level image properties by using fuzzy quantifiers [15]. Image regions which match the target content requirements can then be used to assess any other specifications (shape, size, colour) which appear in the same requirement subtree within the query. Groups of regions which are deemed salient with respect to the query can be compared for the purpose of evaluating relations as mentioned above.


Brush size: large

☐ straight line

☐ vertical

☐ horizontal

☑ stripe

☐ right angl

☐ top edge

☐ left edge

☐ right edge

☐ bottom edge

☑ Mean shape colour

Internal walls: 0.0%

Figure 10: **Examples of alternate ICON query interfaces using sketch of classified target content (left) and region properties (right).**

6 Evaluation

6.1 Qualitative and Quantitative Evaluation

Progress in CBIR research remains hampered by a lack of standards for comparative performance evaluation [31, 30]. This is an intrinsic problem due to the extreme ambiguity of visual information with respect to human vs computer interpretations of content and the strong dependence of relevance assessment upon the particular feature sets and query methods implemented by a given system. There are no publicly available image sets with associated ground truth data at the different levels of granularity required to do justice to different retrieval approaches, nor are there any standard sets of queries and manually ranked results which could easily be translated to the different formats and conventions adopted by different CBIR systems. Furthermore, there are no usable techniques for assessing important yet elusive usability criteria relating to the query interface as discussed in 1.1. Real-world users (rarely addressed in the CBIR research literature) would be primarily interested in the ease with which they could formulate effective queries in a particular system to solve their search requirements with minimal effort for their chosen data set. Even if large scale standardised test sets and sample queries were available to the CBIR community, results derived from them might not be of much use in predicting performance on real-world retrieval tasks.

However, meaningful evaluation of retrieval methods is possible if carried out for a set of well specified retrieval tasks using the same underlying content representation and image database. We have assessed the performance of the OQUEL language in terms of its utility as a query tool both in terms of user effort and query performance. The ICON system has been in use at our research lab and was demonstrated at conferences such as ICCV2001 and CVPR2001. Qualitatively speaking, users find that the OQUEL language provides a more natural and efficient mechanism for

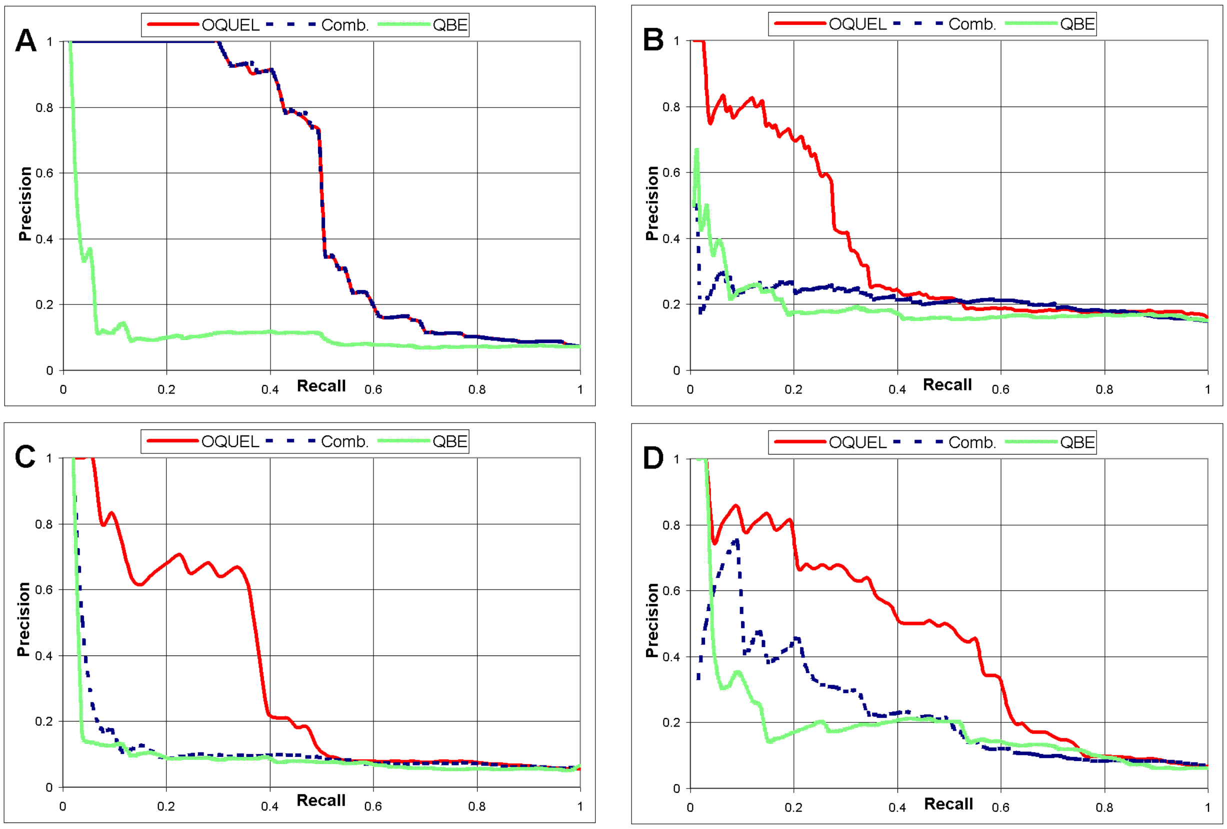


Figure 11: *Plots of relative percentages for precision versus recall for the 4 retrieval experiments.*

content-based querying than the other query methods present in ICON.

6.2 Experimental Method

While most evaluation of CBIR systems is performed on commercial image collections such as the Corel image sets, their usefulness is limited by the fact that they consist of very high quality photographic images and that the associated ground truth (category labels such as “China”, “Mountains”, “Food”) are frequently too high level and sparse to be of use in performance analysis [30]. We therefore chose a set of images consisting of around 670 Corel images augmented with 412 amateur digital pictures of highly variable quality and content. Manual relevance assessments in terms of relevant vs. non-relevant were carried out for all 1082 images over the test queries described below. In the case of the four test queries A–D below, the number of relevant images was 77, 158, 53, and 67 respectively.

In order to quantify the performance of our implementation of the OQUEL language in light of the inherent difficulties of CBIR evaluation, we focussed on contrasting its utility as a retrieval tool compared with the other query modalities present in the ICON system. They are:

- *Query-by-example*: A set of weighted sample images (both positive and negative examples). Comparisons are performed on the basis of metrics such as a pair-wise best region match criterion and a classification pyramid distance measure.

- *User drawn sketch*: Desired image content composed by means of a sketch based query composition tool which uses a visual thesaurus of target image content corresponding to the set of visual categories.

- *Feature range or predicate*: Constraints on visual appearance features (colour, shape, texture) derived from the region segmentation.

The user may assign different weights to the various elements that comprise a query and can choose from a set of similarity metrics to specify the emphasis that is to be placed on the absolute position of target content within images and overall compositional aspects. All of the query components have access to the same precomputed image representation as described in 4.1.

6.3 Test Queries

We chose four test queries which have the following expressions in the OQUEL language:

- *Query A* “bright red and stripy”
- *Query B* “people in centre”
- *Query C* “some water in the bottom half which is surrounded by trees and grass, size at least 10”%
- *Query D* “winter”

These are not meant to constitute a representative sample over all possible image queries (no such sample exists) but to illustrate performance and user search effort for conceptually different retrieval needs expressed at different levels of description. For each OQUEL query we created a further two queries expressed using the other search facilities of the ICON system:

- Composite query: a query which may combine a sketch with feature constraints as appropriate to yield best performance in reasonable time.
- Query-by-example: the single image maximising the normalised average rank metric was chosen as the query. This type of query is commonly used to assess baseline performance.

Figures 6, 7, 8, and 9 show the four OQUEL queries and their search results over the collection. Figure 10 depicts examples of alternate queries consisting of a combination of low level attributes and user drawn sketch.

6.4 Results

To quantify performance, precision vs. recall was computed using the ground truth images for each test query as shown in figure 11. For each OQUEL query we also show the results for a combined query (“Comb.”) and a query-by-example (“QBE”) designed and optimised to meet the same user search requirement. It can be seen that OQUEL queries yield better results, especially for the top ranked images. In the case of query A, results are essentially the same as those for a query consisting of feature predicates for the region properties “stripy” and “red”. In general OQUEL queries are more robust to errors in the segmentation and region classification due to their ontological structure. Query-by-example in particular is usually insufficient to express more advanced concepts relating to spatial composition, feature in variances, or object level constraints.

As recommended in [31], we also computed the *normalised average rank* which is a useful stable measure of relative performance in CBIR:

$$Rank^{\sim} = \frac{1}{NN_{rel}}\{\sum_{i=1}^{N_{rel}} R_i - \frac{N_{rel}\{N_{rel}-1\}}{2}\} \qquad (4)$$

where R_i is the rank at which the i’th relevant image is retrieved, N_{rel} the number of relevant images, N the total number of images in the collection. The value of $Rank^{\sim}$ ranges from 0 to 1 where 0 indicates perfect retrieval.

Query	$Rank^{\sim}$
A - OQUEL	0.2185
A - Comb.	0.2184
A - QBE	0.3992
B - OQUEL	0.2924
B - Comb.	0.3081
B - QBE	0.3693
C - OQUEL	0.2637
C - Comb.	0.3159
C - QBE	0.3530
D - OQUEL	0.1944
D - Comb.	0.2582
D - QBE	0.2586

Comparisons with other query composition and retrieval paradigms implemented in ICON (sketch, sample images, property thresholds) therefore show that the OQUEL query language constitutes a more efficient and flexible retrieval tool. Few prior interpretative constraints are imposed and relevance assessments are carried out solely on the basis of the syntax and semantics of the query itself. Text queries have also generally proven to be more efficient to evaluate since one only needs to analyse those aspects of the image content representation which are relevant to nodes in the corresponding syntax tree and because of various possible optimisations in the order of evaluation to quickly rule out non-relevant images. Although the current system does not use an inverted file as its index, query evaluation took no more than 100ms for the test queries.

7 Conclusions

7.1 Discussion

As explained above, one of the primary advantages of the proposed language-based query paradigm for CBIR is the ability to leave the problem of initial domain selection to the user. The retrieval process operates on a description of desired image content expressed according to an ontological hierarchy defined by the language and relates this at retrieval time to the available image content representation. Domain knowledge therefore exists at three levels: the structure and content of the user query, the ontology underlying the query language, and the retrieval mechanism which parses the user query and assesses image relevance. User queries may be quite high-level and employ general terms, thus placing the burden of finding feature combinations which discriminate relevant from non-relevant images on the ontology and the interpreter. Richer, more specific queries narrow down the retrieval focus. One can therefore offset user composition effort and the need for greater language and parser complexity depending on the relative costs involved in a real world CBIR context.

The current implementation does not constitute an exhaustive means of mapping retrieval requirements and relating them to images. Nor does the OQUEL language come close to embodying the full richness of a natural language specification of concepts relating to properties of photographic images. However, the current system does show that it is possible to utilise an ontological language framework to fuse different individually weak and ambiguous sources of image information and content representation in a way which improves retrieval performance and usability of the system. Clearly there remain scalability issues as additional classifiers will need to be added to improve the representational capacity of the query language. However, the notion of ontology based languages provides a powerful tool for extending retrieval systems by adding task and domain specific concept hierarchies at different levels of semantic granularity.

7.2 Summary and Future Outlook

Query composition is a relatively ill-understood part of research into CBIR and clearly merits greater attention if image retrieval systems are to enter the mainstream. Most systems for content based image retrieval offer query composition facilities based on examples, sketches, feature predicates, structured database queries, or keyword annotation. Compared to document retrieval using text queries, user search effort remains significantly higher, both in terms of initial query formulation and the need for relevance feedback.

This paper argues that query languages provide a flexible way of dealing with problems commonly encountered in CBIR such as ambiguity of image content and user intention and the semantic gap which exists between user and system notions of relevance. By basing such a language on an extensible ontology, one can explicitly state ontological commitments about categories, objects, attributes, and relations without having to pre-define any particular method of query evaluation or image interpretation. A central theme of the paper is the distinction between query description and image description languages and the power of a formally specifiable language featuring syntax and semantics in order to capture meaning in images relative to a query. The combination of individually weak and ambiguous clues to determine object presence and estimate overall probability of relevance builds on recent approaches to robust object recognition and can be seen as an attempt at extending the success of indicative methods for content representation in the field of text retrieval.

We present *OQUEL* as an example of such a language. It is a novel query description language which works on the basis of short text queries describing the user's retrieval needs and does not rely on prior annotation of images. Query sentences can represent abstract and arbitrarily complex retrieval requirements at multiple levels and integrate multiple sources of evidence. The query language itself can be extended to represent customised ontologies defined on the basis of existing terms. An implementation of OQUEL for the ICON system demonstrates that efficient retrieval of general photographic images is possible through the use of short OQUEL queries consisting of natural language words and a simple syntax.

The use of more sophisticated natural language processing techniques would ease the current grammatical restrictions imposed by the syntax and allow statistical interpretation of more free-form query sentences consisting of words from an extended vocabulary. Perhaps most importantly, ongoing efforts aim to acquire the weighting of the Bayesian inference nets used in scene and object recognition using a training corpus and prior probabilities for the visual categories. The goal is to reduce the need for pre-wired knowledge such as "an image containing regions of snow and ice is more likely to depict a winter scene". An approach such as [11] paired with the structural Expectation Maximisation method might provide a means of automatically acquiring new high level terms and their inference networks. The automated discovery of domain and general purpose ontologies together with the means of relating these to lower level evidence is an important challenge for data mining and machine learning research.

Acknowledgements

The authors would like to acknowledge directional guidance and support from AT&T Laboratories Research and financial assistance from the Royal Commission for the Exhibition of 1851.

References

[1] A. Abella and J. Kender. From pictures to words: Generating locative descriptions of objects in an image. *ARPA94*, pages II:909–918, 1994.

[2] M. Aiello, C. Areces, and M. de Rijke. Spatial reasoning for image retrieval. In *Proc. International Workshop on Description Logics*, 1999.

[3] S. Bechhofer and C. Goble. Description logics and multimedia—applying lessons learnt from the Galen project. In *Proc. Workshop on Knowledge Representation for Interactive Multimedia Systems*, 1996.

[4] A. Bobick and W. Richards. Classifying objects from visual information. Technical report, MIT AI Lab, 1986.

[5] N. Campbell, W. Mackeown, B. Thomas, and T. Troscianko. Interpreting image databases by region classification. *Pattern Recognition (Special Edition on Image Databases)*, 30(4):555–563, April 1997.

[6] C. Carson, S. Belongie, H. Greenspan, and J. Malik. Region-based image querying. In *Proc. IEEE Workshop on Content-based Access of Image and Video Libraries*, 1997.

[7] S. Chang, W. Chen, and H. Sundaram. Semantic visual templates: Linking visual features to semantics. In

Proc. Workshop on Content Based Video Search and Retrieval, 1998.

[8] T.-S. Chua, K.-C. Teo, B.-C. Ooi, and K.-L. Tan. Using domain knowledge in querying image databases. In *International Conference on Multimedia Modeling*, 1996.

[9] I. Cox, M. Minka, T. Minka, T. Papathomas, and P. Yianilos. The bayesian image retrieval system, PicHunter. *IEEE Transactions on Image Processing*, 9(1):20–37, 2000.

[10] M. Dobie, T. Robert, D. Joyce, M. Weal, P. Lewis, and W. Hall. A flexible architecture for content and concept based multimedia information exploration. In *Proc. Second UK Conference on Image Retrieval*, 1999.

[11] P. Duygulu, K. Barnard, J.F.H. De Freitas, and D.A. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Proc. European Conference on Computer Vision*, 2002.

[12] A. Evans, N. Thacker, and J. Mayhew. The use of geometric histograms for model-based object recognition. In *Proc. British Machine Vision Conference*, 1993.

[13] D. Fensel, I. Horrocks, F. van Harmelen, S. Decker, M. Erdmann, and M. Klein. OIL in a nutshell. In *Knowledge Acquisition, Modeling and Management*, pages 1–16, 2000.

[14] C. Fung and K. Loe. A new approach for image classification and retrieval. In *Proc. ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 301–302. ACM, 1999.

[15] I. Glöckner and A. Knoll. Fuzzy quantifiers for processing natural-language queries in content-based multimedia. Technical Report TR97-05, Faculty of Technology, University of Bielefeld, Germany, 1997.

[16] T. Gruber. Towards Principles for the Design of Ontologies Used for Knowledge Sharing. *Formal Ontology in Conceptual Analysis and Knowledge Representation*, 1993.

[17] V. Gudivada and V. Raghavan. Content-based image retrieval systems. *IEEE Computer*, 28(9):18–22, 1995.

[18] M. Hacid and C. Rigotti. Representing and Reasoning on Conceptual Queries Over Image Databases. In *Proc. of the Eleventh International Symposium on Methodologies for Intelligent Systems*, LNCS 1609, pages 340–348. Springer, 1999.

[19] C. Hsu, W. Chu, and R. Taira. A knowledge-based approach for retrieving images by content. *Knowledge and Data Engineering*, 8(4):522–532, 1996.

[20] W. Hu. An overview of the world wide web search technologies. In *Proc. SCI2001 – 5th World Multiconference on System, Cybernetics and Informatics*, 2001.

[21] ICON System, AT&T Laboratories Cambridge. `www.uk.research.att.com/permm/icon.html`.

[22] A. Jaimes and S. Chang. Model-based classification of visual information for content-based retrieval. In *Proc. SPIE Conference on Storage and Retrieval for Image and Video Databases*, 1999.

[23] T. Kato, T. Kurita, N. Otsu, and K. Hirata. A sketch retrieval method for full color image database query by visual example. In *Proc. Int. Conference on Pattern Recognition*, pages I:530–533, 1992.

[24] P. Kelly, T. Cannon, and D. Hush. Query by image example: The comparison algorithm for navigating digital image databases (CANDID) approach. In *Storage and Retrieval for Image and Video Databases (SPIE)*, pages 238–248, 1995.

[25] M. Lalmas. *Applications of Uncertainty Formalisms*, chapter Information retrieval and Dempster-Shafer's theory of evidence, pages 157–177. Springer, 1998.

[26] J. Lim. Learnable visual keywords for image classification. In *Proc. ACM International Conference on Digital Libraries*, 1999.

[27] G. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller. Introduction to Wordnet: an on-line lexical database. *International Journal of Lexicography*, 3:235–244, 1990.

[28] T. Mills, D. Pye, D. Sinclair, and K. Wood. Shoebox: A digital photo management system. Technical report, AT&T Laboratories Cambridge, 1999.

[29] A. Mojsilovic, J. Gomes, and B. Rogowitz. Isee: Perceptual features for image library navigation. In *Proc. 2002 SPIE Human Vision and Electronic Imaging*, 2002.

[30] H. Mueller, S. Marchand-Maillet, and T. Pun. The truth about corel - evaluation in image retrieval. In *Proc. Conf. on Image and Video Retrieval*, LNCS 2383, pages 38–50. Springer, 2002.

[31] H. Mueller, W. Mueller, D. McG Squire, S. Marchand-Maillet, and T. Pun. Performance evaluation in content-based image retrieval: Overview and proposals. *Pattern Recognition Letters*, 22(5):593–601, 2001.

[32] S. Nepal, M. Ramakrishna, and J. Thom. A fuzzy object query language (foql) for image databases. In *Sixth International Conference on Database Systems for Advanced Applications*, 1999.

[33] W. Niblack. The qbic project: querying images by color, texture and shape. Technical report, IBM Research Report RJ-9203, 1993.

[34] V. Ogle and M. Stonebraker. Chabot: Retrieval from a relational database of images. *IEEE Computer*, 28(9):40–48, 1995.

[35] S. Parsons and A. Hunter. *Applications of Uncertainty Formalisms*, chapter A review of uncertainty handling formalisms, pages 8–37. Springer, 1998.

[36] G.M. Petrakis. Design and evaluation of spatial similarity approaches for image retrieval. *Image and Vision Computing*, 20:59–76, 2002.

[37] K. Rodden. How do people organise their photographs? In *BCS IRSG 21st Annual Colloquium on Information Retrieval Research*, 1999.

[38] N. Roussoupolos, C. Falautsos, and T. Sellis. An efficient pictorial database system for psql. *IEEE Transactions on Software Engineering*, 14(5):639–659, 1988.

[39] W. Rucklidge. Locating ojects using the hausdorff distance. In *Proc. International Conference on Computer Vision*, 1995.

[40] Y. Rui, T. Huang, and S. Chang. Image retrieval: Current techniques, promising directions and open issues. *Journal of Visual Communication and Image Representation*, 10:39–62, 1999.

[41] S. Santini, A. Gupta, and R. Jain. Emergent semantics through interaction in image databases. *Knowledge and Data Engineering*, 13(3):337–351, 2001.

[42] H. Shen and K. Ooi, B. Tan. Finding semantically related images in the WWW. In *Proc. ACM Multimedia*, pages 491–492, 2000.

[43] D. Sinclair. Voronoi seeded colour image segmentation. Technical Report TR99-04, AT&T Laboratories Cambridge, 1999.

[44] D. Sinclair. Smooth region structure: folds, domes, bowls, ridges, valleys and slopes. In *Proc. Conference on Computer Vision and Pattern Recognition*, pages 389–394. IEEE Comput. Soc. Press, 2000.

[45] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, 2000.

[46] J. Smith and S. Chang. Image and video search engine for the world wide web. In *Storage and Retrieval for Image and Video Databases (SPIE)*, pages 84–95, 1997.

[47] K. Sparck Jones. Information retrieval and artificial intelligence. *Artificial Intelligence*, 114:257–281, 1999.

[48] K. Tieu and P. Viola. Boosting image retrieval. In *Proc. International Conference on Computer Vision*, 2000.

[49] C.P. Town and D.A. Sinclair. Content based image retrieval using semantic visual categories. Technical Report MV01-211, Society for Manufacturing Engineers, 2001.

[50] N. Vasconcelos and A. Lippman. A bayesian framework for content-based indexing and retrieval. In *Proc. Conference on Computer Vision and Pattern Recognition*, 1998.

[51] L. Wenyin, S. Dumais, Y. Sun, H. Zhang, M. Czerwinski, and B. Field. Semi-automatic image annotation. In *Proc. Interact2001 Conference on Human Computer Interaction*, 2001.

[52] M. Wood, N. Campbell, and B. Thomas. Iterative refinement by relevance feedback in content-based digital image retrieval. In *Proc. ACM Multimedia 98*, pages 13–17, 1998.