Ontological Query Language for Content Based Image Retrieval

AT&T Labs Cambridge technical report 2001.1

Chris Town David Sinclair

AT&T Laboratories Cambridge
24a Trumpington Street
Cambridge CB2 1QA, England

Abstract

This paper discusses the design and implementation of the oquel query language for content based image retrieval. 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", "buildings", etc. on the basis of existing language constructs.

Images are retrieved by deriving an abstract syntax tree from of a textual user query and probabilistically evaluating it by comparing the composition and perceptual properties of salient image regions in light of the query. The matching process utilises automatically extracted image segmentation and classification information and can incorporate any other feature detection mechanisms (such as face recognisers) or context-dependent knowledge available at the time the query is processed.

1 Introduction

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 to retrieve documents from vast online collections, it seems natural to extend language based querying to multimedia data. Content based image retrieval (CBIR) on the basis of short query sentences is likely to prove more efficient and intuitive than alternative query composition schemes such as iterative search-by-example and user sketches which are employed by most current systems.

However, 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. Formal query languages such as extensions of SQL [9] are limited in their expressive power and extensibility and require a certain level of user experience and sophistication. Other efforts which rely on the pre-computation of object-relational graph structures ([8], [6]) are computationally expensive and may require complex queries to yield the expected results.

In order to address the challenges mentioned above while keeping user search overheads at a minimum, we have developed the *oquel* query description language. It provides an extensible language framework based on a context free grammar and a base vocabulary. 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 differs in a number of respects from related attempts at using language or semantic graphs to facilitate content based access to image collections ([3], [8], [7], [4]). It is portable to other image content description 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 allows the definition of customised ontologies of objects and abstract relations. The process of assessing image relevance can be made dynamic in the sense that the way in which elements of a query are evaluated depends on the query as a whole (information flows both up and down) and any domain specific information with respect to the ontological makeup of the query which may be available at the time it is processed.

This paper discusses the basic design and structure of the oquel language and illustrates the retrieval process through sample queries. The discussion is based on an implementation of the language for the *ICON* content-based image retrieval system.

2 Language design and structure

2.1 Overview

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. Ongoing work seeks to extend the language core discussed here to provide more advanced programmatic constructs offering capabilities familiar from database query languages and to enable autonomous learning of new language concepts.

Oquel queries (sentences) are prescriptive rather than descriptive, i.e. the focus is on making it easy to formulate desired image characteristics as concisely as possible. It is therefore neither necessary nor desirable to provide an exhaustive description of the visual features and semantic content of particular images. Instead a query represents only as much information as is required to discriminate relevant from non-relevant images.

2.2 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"). The latter includes a distinction between singular and plural, 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.

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 an LALR(1) grammar specification. This includes a thesaurus of several hundred natural language words, phrases, and abbreviations (e.g. "!" for "not") which are recognised as tokens.

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).

$$
\begin{aligned}
G : \{ & \\
S &\rightarrow R \\
R &\rightarrow \mathit{modifier}?\ (\mathit{metacategory} \mid SB \mid BR) \\
& \quad \mid \mathit{not}?\ R\ (CB\ R)? \\
BR &\rightarrow SB\ \mathit{binaryrelation}\ SB \\
SB &\rightarrow (CS \mid PS) +\ LS * \\
CS &\rightarrow \mathit{visualcategory} \mid \mathit{semanticcategory} \mid \\
& \quad \mathit{not}?\ CS\ (CB\ CS)? \\
LS &\rightarrow \mathit{location} \mid \mathit{not}?\ LS\ (CB\ LS)? \\
PS &\rightarrow \mathit{shapedescriptor} \mid \mathit{colourdescriptor} \mid \\
& \quad \mathit{sizedescriptor} \mid \mathit{not}?\ PS\ (CB\ PS)? \\
CB &\rightarrow \mathit{and} \mid \mathit{or} \mid \mathit{xor}; \\
\} &
\end{aligned}
$$

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
CB: binary (fuzzy) logical connective

Tokens (terminals) belong to the following sets:
modifier: Quantifiers such as "a lot of", "none", "as much as possible".
metacategory: Scene descriptors which apply over the entire image, e.g. countryside, city, indoors.
binaryrelation: To specify 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..
visualcategory: Categories of stuff, e.g. water, skin.

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", "angled".
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".

2.3 Example sentences

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

> some sky which is close to trees in upper corner, size at least 20%
>
> [indoors] or [outdoors] & [people]
>
> [some green or vividly coloured vegetation in the centre] which is of similar size as [clouds or blue sky at the top]
>
> artificial objects, smooth and polygonal

3 Query evaluation and retrieval

This section illustrates the oquel retrieval process as implemented in the ICON (Image Content Organisation and Navigation, [2]) system. This combines a cross-platform Java user interface with image processing and content analysis functionality to facilitate automated organisation and retrieval of large heterogeneous image sets based on both meta data and visual content.

3.1 Content representation

ICON extracts various types of content descriptors and meta data from images (see [11]). The following are currently used when evaluating oquel text queries:

Image segmentation: Images are segmented into non-overlapping regions and sets of properties are computed for each region (see [10]). A mask gives the absolute location of each region.

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" such as grass, sky, and skin.

Region graph: graph of the relative spatial relationships of the regions (adjacency, distance, joint boundary, and containment).
Grid pyramid: for each visual category, proportion of image content which has been positively classified (as computed by the region labelling) at different regular grid spacings (1x1, image fifths, 8x8).

The choice of visual categories 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. Through the relationship graph representation of regions we can make the matching of clusters of regions invariant with respect to displacement and rotation, whereas the grid pyramid representation caters for a 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 [5].

3.2 Query evaluation and retrieval

Images are retrieved by evaluating an abstract syntax tree (AST) derived from the user query to compute a probability of relevance for each image. In the first stage, the AST is parsed depth first and the leaf nodes are evaluated in light of their predecessors and siblings. Information then propagates back up the tree until we are 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 whether people are present in the image. Ongoing efforts aim to acquire the weighting of the Bayesian inference nets 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".

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)). 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

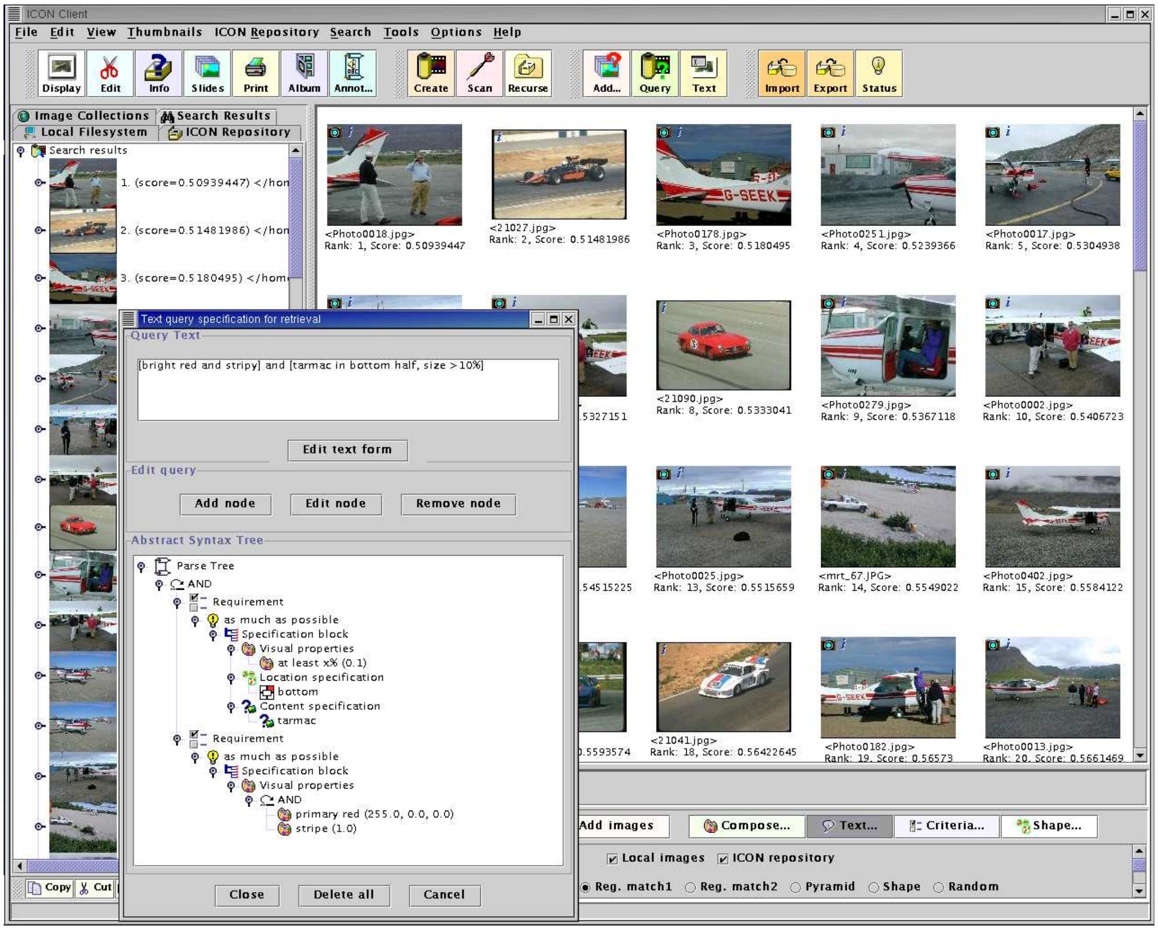

Figure 1: *Search results for the text query "[bright red and stripy] and [tarmac in bottom half, size >10%]".*

query. Groups of regions which are deemed salient for the purposes of the query can also be compared using the binary relations mentioned above.

Comparisons with other query composition and retrieval paradigms implemented in ICON (sketch, sample images, property thresholds) show that the oquel query language constitutes a much 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 we only need to analyse those aspects of the image content representation which are relevant to nodes in the corresponding AST an because of various possible optimisations in the order of evaluation.

Figures 1 and 2 show simple oquel queries and search results.

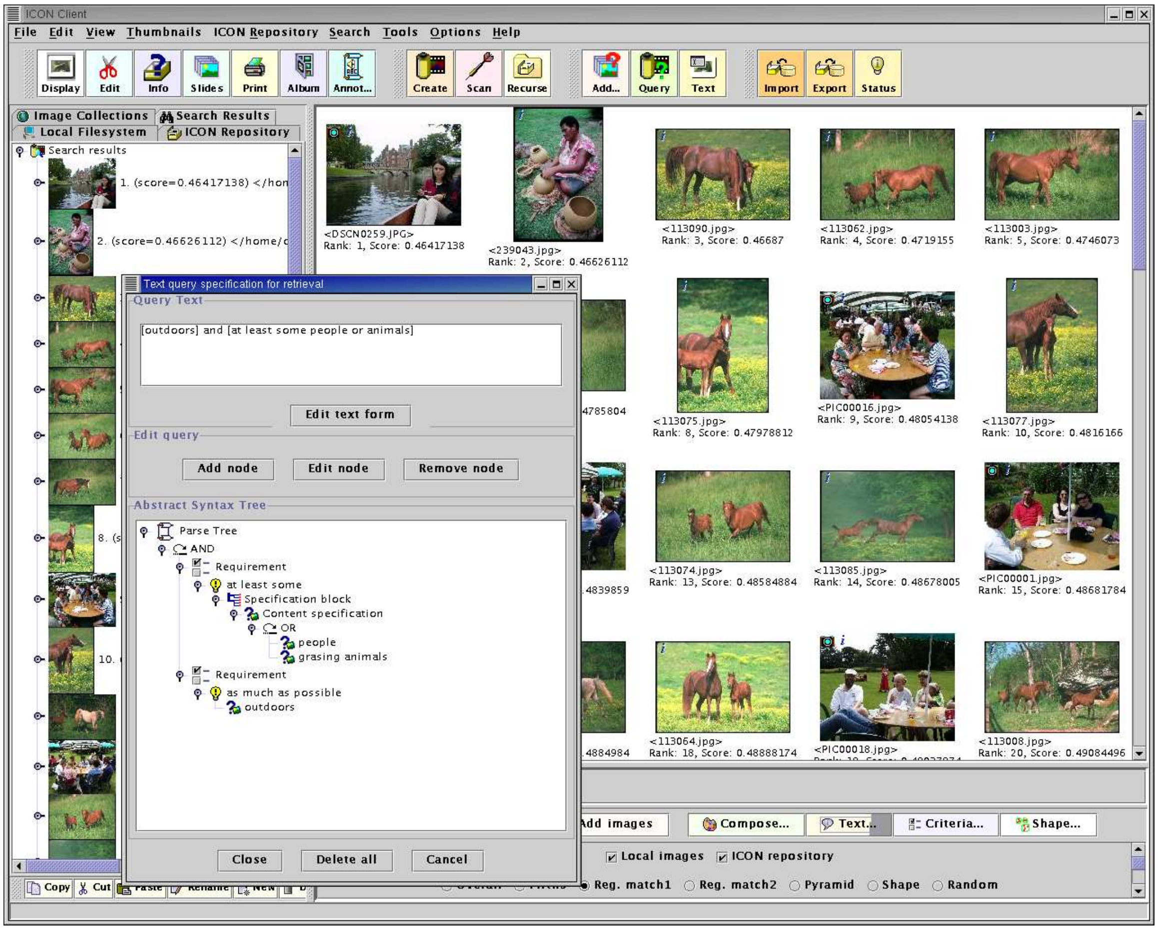

Figure 2: *Search results for the text query "[outdoors] and [at least some people or animals]".*

4 Summary and outlook

Most systems for content based image retrieval (CBIR) offer query composition facilities based on examples, sketches, structured database queries, or keywords. Compared to document retrieval using text queries, user search effort remains significantly higher, both in terms of initial query formulation and relevance feedback. This paper discusses oquel, a novel query description language for CBIR which works on the basis of short text queries describing the user's retrieval needs and does not rely on prior annotation of images. The language 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. Further work to enrich the language for the purposes of retrieval from professional image libraries is in progress.

Acknowledgements

The authors would like to acknowledge directional guidance and encouragement from Professor Andy Hopper.

References

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

[2] AT&T Laboratories Cambridge. ICON System. In `http://www.uk.research.att.com/permm/icon.html`.

[3] N. S. Chang and K. S. Fu. Picture query languages for pictorial data-base systems. *IEEE COMPUTER*, 14, 11:23–33, 1981.

[4] W.W. Chu, C. Hsu, A.F. Cardenas, and R.K. Taira. Knowledge-based image retrieval with spatial and temporal constructs. *IEEE Transactions on Knowledge and Data Engineering*, 10, 6, 1998.

[5] D. Forsyth, J. Malik, M. Fleck, and J. Ponce. Primitives, perceptual organization and object recognition. Technical report, Computer Science Division, University of California at Berkeley, 1997.

[6] T. Hermes, C. Klauck, J. Kreys, and J. Zhang. Image retrieval for information systems. In *Storage and Retrieval for Image and Video Databases (SPIE)*, pages 394–405, 1995.

[7] M. Mechkour, C. Berrut, and Y. Chiaramella. Using conceptual graph framework for image retrieval, 1995.

[8] E.G.M. Petrakis and C. Faloutsos. Similarity searching in large image databases. Technical Report CS-TR-3388, College Park, MD, USA, 1994.

[9] N. Roussoupolos, C. Falautsos, and T. Sellis. An efficient pictorial database system for psql, 1988.

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

[11] C.P. Town and D. Sinclair. Content based image retrieval using semantic visual categories. Technical Report TR2000-14, AT&T Laboratories Cambridge, 2000.