Machine Vision and Applications manuscript No.
(will be inserted by the editor)

Christopher Town

Ontological Inference for Image and Video Analysis

Received: date / Accepted: date

Abstract This paper presents an approach to designing and implementing extensible computational models for perceiving systems based on a knowledge-driven joint inference approach. These models can integrate different sources of information both horizontally (multi-modal and temporal fusion) and vertically (bottom-up, top-down) by incorporating prior hierarchical knowledge expressed as an extensible ontology.

Two implementations of this approach are presented. The first consists of a content based image retrieval system which allows users to search image databases using an ontological query language. Queries are parsed using a probabilistic grammar and Bayesian networks to map high level concepts onto low level image descriptors, thereby bridging the "semantic gap" between users and the retrieval system. The second application extends the notion of ontological languages to video event detection. It is shown how effective high-level state and event recognition mechanisms can be learned from a set of annotated training sequences by incorporating syntactic and semantic constraints represented by an ontology.

Keywords Ontologies · Perceptual inference · Content-based image retrieval · Video analysis · Knowledge-based computer vision

1 Introduction

Visual information is inherently ambiguous and semantically impoverished. There consequently exists a wide semantic gap between human interpretations of image and video data and that currently derivable by means of a computer. This paper demonstrates how this gap can be narrowed by means of ontologies. Ontology is the theory of objects in terms of the criteria which allow one to distinguish between different types of objects and their relationships, dependencies, and properties. Ontologies encode the relational structure of concepts which one can use to describe and reason about aspects of the world. This makes them eminently suitable to many problems in computer vision which require prior knowledge to be modelled and utilised in both a descriptive and prescriptive capacity.

In this paper, terms in the ontology are grounded in the data and therefore carry meaning directly related to the appearance of real world objects. Tasks such as image retrieval and automated visual surveillance can then be carried out by processing sentences in a visual language defined over the ontology. Such sentences are not purely symbolic since they retain a linkage between the symbol and signal levels. They can therefore serve as a computational vehicle for active knowledge representation which permits incremental refinement of alternate hypotheses through the fusion of multiple sources of information and goal-directed feedback. A visual language can also serve as an important mechanism for attentional control by constraining the range of plausible feature configurations that need to be considered when performing a visual task such as recognition. Processing may then be performed selectively in response to queries formulated in terms of the structure of the domain, i.e. relating high-level symbolic representations to extracted visual and temporal features in the signal. By basing such a language on an ontology one can capture both concrete and abstract relationships between salient visual properties. Since the language is used to express queries and candidate hypotheses 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 or video. Instead, only those image aspects which are of value given a particular task are evaluated and

University of Cambridge Computer Laboratory, 15 JJ Thomson Avenue, Cambridge CB3 0FD, UK
Tel: +44 (0)1223 763500
Fax: +44 (0)1223 334678
E-mail: cpt23@cam.ac.uk

evaluation may stop as soon as the appropriate top level symbol sequence has been generated.

This approach is broadly motivated by two notions of how visual information processing may be achieved in biological and artificial systems. Firstly, vision can be posed as knowledge-driven probabilistic inference. Mathematical techniques for deductive and inductive reasoning can then be applied to deal with two key problems that make vision difficult, namely complexity and uncertainty. Recognition is thus posed as a joint inference problem relying on the integration of multiple (weak) clues to disambiguate and combine evidence in the most suitable context as defined by the top level model structure.

Secondly, vision may be regarded as closely related to (and perhaps an evolutionary precursor of) language processing. In both cases one ultimately seeks to find symbolic interpretations of underlying signal data. Such an analysis needs to incorporate a notion of the syntax and semantics that is seen as governing the domain of interest so that the most likely explanation of the observed data can be found. The general idea is that recognising an object or event requires one to relate loosely defined symbolic representations of concepts to concrete instances of the referenced object or behaviour pattern. This is best approached in a hierarchical manner by associating individual parts at each level of the hierarchy according to rules governing which configurations of the underlying primitives give rise to meaningful patterns at the higher semantic level. Thus syntactic rules can be used to drive the recognition of compound objects or events based on the detection of individual components corresponding to detected features in time and space. Visual analysis then amounts to parsing a stream of basic symbols according to prior probabilities to find the most likely interpretation of the observed data in light of the top-level starting symbols in order to establish correspondence between numerical and symbolic descriptions of information.

This paper presents two concrete implementations of the approach discussed above which demonstrate its utility for solving relevant research problems.

2 Related work

2.1 Visual recognition as perceptual inference

An increasing number of research efforts in medium and high level video analysis can be viewed as following the emerging trend that object recognition and the recognition of temporal events are best approached in terms of generalised language processing which attempts a machine translation [15] from information in the visual domain to symbols and strings composed of predicates, objects, and relations. Many state-of-the-art recognition systems therefore explicitly or implicitly employ a probabilistic grammar which defines the syntactic rules which can be used to recognise compound objects or events based This idea has a relatively long heritage in syntactic approaches to pattern recognition ([66],[7]) but interest has been revived recently in the video analysis community following the popularity and success of probabilistic methods such as Hidden Markov models (HMM) and related approaches adopted from the speech and language processing community.

While this approach has shown great promise for applications ranging from image retrieval to face detection to visual surveillance, a number of problems remain to be solved. The nature of visual information poses hard challenges which hinder the extent to which mechanisms such as Hidden Markov models and stochastic parsing techniques popular in the speech and language processing community can be applied to information extraction from images and video. Consequently there remains some lack of understanding as to which mechanisms are most suitable for representing and utilising the syntactic and semantic structure of visual information and how such frameworks can best be instantiated. The role of machine learning in computer vision continues to grow and recently there has been a very strong trend towards using Bayesian techniques for learning and inference, especially factorised graphical probabilistic models [27] such as Dynamic Belief networks (DBN). While finding the right structural assumptions and prior probability distributions needed to instantiate such models requires some domain specific insights, Bayesian graphs generally offer greater conceptual transparency than e.g. neural network models since the underlying causal links and prior beliefs are made more explicit. The recent development of various approximation schemes based on iterative parameter variation or stochastic sampling for inference and learning have allowed researchers to construct probabilistic models of sufficient size to integrate multiple sources of information and model complex multi-modal state distributions. Recognition can then be posed as a joint inference problem relying on the integration of multiple (weak) clues to disambiguate and combine evidence in the most suitable context as defined by the top level model structure.

As illustrated by [13] and [60], concurrent probabilistic integration of multiple complementary and redundant cues can greatly increase the robustness of multi-hypothesis tracking. In [54] tracking of a person's head and hands is performed using a Bayesian Belief network which deduces the body part positions by fusing colour, motion and coarse intensity measurements with context dependent semantics. Later work by the same authors [55] again shows how multiple sources of evidence (split into necessary and contingent modalities) for object position and identity can be fused in a continuous Bayesian framework together with an observation exclusion mechanism.

An approach to visual tracking based on co-inference of multiple modalities is also presented in [69] which describes an sequential Monte Carlo approach to co-infer target object colour, shape, and position. In [9] a joint probability data association filter (JPDAF) is used to compute the HMM's transition probabilities by taking into account correlations between temporally and spatially related measurements. [22] presents a method for recognising video events using a tracking framework and Bayesian networks based on shape and trajectory information. Composite events are analysed using a semi-hidden Markov Model exhibiting better performance than standard HMMs on noisy sequences.

2.2 Linking language to visual data

In the area of still image descriptions, Abella and Kender ([2,1]) demonstrated a method for generating path and location descriptions from images such as maps and specialist medical images. Spatial prepositions are represented using predicates in fuzzy logic and combined with prior and task specific knowledge to generate natural language expressions concerning spaces and locations.

[68,59] describe a system that uses Bayesian networks to integrate verbal descriptions of objects (colour, size, type) and spatial relationships in a scene with features and classifications resulting from image processing. The network is generated from the two forms of representation by matching object properties and relations extracted from the visual and speech processing.

In a similar vein, [52,53] uses machine learning to establish correspondences between objects in a scene and natural language descriptions of them. Words in the vocabulary are grounded in a feature space by computing the KL-divergence of the probability distribution for a given word conditioned on a particular feature set and the unconditioned distribution. Co-occurrence frequencies and word bigram statistics are used to learn semantic associations of adjectives (including spatial relationships) and noun order respectively. The training process relies on human descriptions of designated objects. Perhaps a larger corpus of such data would make an approach such as [4] feasible which matches still image annotations with region properties using hierarchical clustering and EM.

Learning associations between visual keywords and image properties is of particular interest for content-based image retrieval [50,34,71,63] where keyword associations can be acquired using a variety of supervised (e.g. neural network) and unsupervised (e.g. latent semantic analysis) learning schemes. These methods are generally restricted to fairly low-level properties and descriptors with limited semantic content. Such information can also be acquired dynamically from user input [26] whereby a user defines visual object models via an object-definition hierarchy (region, perceptual-area, object part, and object).

Recent work [5,3] has shown some promising results with methods using hierarchical clustering to learn the joint probability distribution of image segment features and associated text, including relatively abstract descriptions of artwork. This uses a generative hierarchical method for EM (Expectation Maximisation, [51]) based learning of the semantic associations between clustered keywords (which are high-level, sparse, and ambiguous denoters of content) and image features (which are semantically poor, visually rich, and concrete) to describe pictures. In order to improve the coherence in the annotations, the system makes use of the WordNet [36] lexical database. This is an interesting approach that is currently being extended to work with natural language image descriptions and more advanced image segmentation and feature extraction.

In [23], information from the WordNet is used to analyse and annotate video sequences. Visual information obtained using face detection, scene classification, and motion tracking is translated into words. These words are then used to generate scene descriptions by performing a search over the semantic relationships present in WordNet. Thus video analysis relies on searching WordNet for concepts jointly supported by video evidence and topic context derived from video transcription.

[29] describes some preliminary work on integrating a novel linguistic question answering method with a video surveillance system. By combining various approaches to temporal reasoning and event recognition from the artificial intelligence community, the authors are proposing a common visual-linguistic representation to allow natural language querying of events occurring in the surveillance footage. A similar problem is considered in [30] which presents a spatio-temporal query language that can be used for analysing traffic surveillance scenarios. The language features unary and binary relations over attributes such as distances, orientations, velocities, and temporal intervals. Queries consisting of trees of such relations are matched to the output of a tracking framework by considering all possible ways of binding tracked objects to leaf nodes in the tree and evaluating relations to assess whether all constraints are matched. In [44] a system for generating verbal descriptions of human movements is presented. The method makes use of a hierarchy of human body parts and actions in order to generate the most plausible and succinct description of movements observed from video sequences.

2.3 Ontologies and hierarchical representations

Many classical methods for representing and matching ontological knowledge in artificial intelligence (description logics, frame-based representations, semantic nets) are coming back into vogue, not least because of the "semantic web" initiative. However, many problems remain when such approaches are applied to highly uncertain

and ambiguous data of the sort that one is confronted with in computer vision and language processing. Much research remains to be done in fusing classical syntactic approaches to knowledge representation with modern factorised probabilistic modelling and inference frameworks.

Early work by Tsotsos [67] presents a mechanism for motion analysis (applied to medical image sequences) based on instantiation of prior knowledge frames represented by semantic networks. The system can maintain multiple hypotheses for the motion descriptors which best describe the movement of objects observed in the sequence. A focus of attention mechanism and a feedback loop featuring competition and reinforcement between different hypotheses are used to rank possible interpretations of a sequence and perform temporal segmentation.

In [10], domain knowledge in the form of a hierarchy of descriptors is used to enhance content-based image retrieval by mapping high-level user queries onto relations over pertinent image annotations and simple visual properties (colour and texture).

In [12], an architecture for perceptual computing is presented which integrates different visual processing routines in the form of a "federation of processes" where bottom-up data is fused with top-down information about the user's context and roles based on an ontology.

The use of such an ontology for information fusion is made more explicit in [31] which uses the DARPA Agent Markup Language (DAML) that was originally developed to facilitate the "semantic web". Their paper considers more of a "toy problem" and doesn't really address problems with description logics of this sort (such as brittleness and the frame problem).

A more robust approach is presented in [43] which describes an event recognition language for video. Events can be hierarchical composites of simpler primitive events defined by various temporal relationships over object movements. Very recently [42], there have been ongoing efforts by the same authors and others to produce a standardised taxonomy for video event recognition consisting of a video event representation language (VERL) and a video event markup language (VEML) for annotation.

[35] uses an ontology of object descriptors to map higher level content-based image retrieval queries onto the outputs of image processing methods. The work seems to be at an early stage and currently relies on several cycles of manual relevance feedback to perform the required concept mappings. Similar work on evaluating conceptual queries expressed as graphs is presented in [16] which uses sub-graph matching to match queries to model templates for video retrieval. In application domains where natural language annotations are available, such as crime scene photographs [46], retrieval can also gain from the extraction of complex syntactic and semantic relationships from image descriptions by means of sophisticated natural language processing.

Ontologies have also been used to extend standardised multimedia annotation frameworks such as MPEG-7 with concept hierarchies [25]. They also play an important role in improving content-based indexing and access to textual documents (e.g. [19], [32]) where they can be used for semantics-based query expansion and document clustering.

3 Proposed approach and methodology

3.1 Overview

We propose a cognitive architectural model for image and video interpretation. It is based on a self-referential probabilistic framework for multi-modal integration of evidence and context-dependent inference given a set of representational or derivational goals. This means that the system maintains an internal representation of its current hypotheses and goals and relates these to available detection and recognition modules. For example, a surveillance application may be concerned with recording and analysing movements of people by using motion estimators, edge trackers, region classifiers, face detectors, shape models, and perceptual grouping operators. The system is capable of maintaining multiple hypotheses at different levels of semantic granularity and can generate a consistent interpretation by evaluating a query expressed in an ontological language. This language gives a probabilistic hierarchical representation incorporating domain specific syntactic and semantic constraints from a visual language specification tailored to a particular application and for the set of available component modules.

From an artificial intelligence point of view, this can be regarded as an approach to the *symbol grounding problem* [20] since sentences in the ontological language have an explicit foundation of evidence in the feature domain, so there is a way of bridging the semantic gap between the signal and symbol level. It also addresses the *frame problem* [14] since there is no need to exhaustively label everything that is going on, one only needs to consider the subset of the state space required to make a decision given a query which implicitly narrows down the focus of attention.

The nature of such queries is task specific. They may either be explicitly stated by the user (e.g. in an image retrieval task) or implicitly derived from some notion of the system's goals. For example, a surveillance task may require the system to register the presence of people who enter a scene, track their movements, and trigger an event if they are seen to behave in a manner deemed "suspicious" such as lingering within the camera's field of view or repeatedly returning to the scene over a short time scale. Internally the system could perform these functions by generating and processing queries of the kind "does the observed region movement correspond to

a person entering the scene?", "has a person of similar appearance been observed recently?", or "is the person emerging from behind the occluding background object the same person who could no longer be tracked a short while ago?". These queries would be phrased in a language which relates them to the corresponding feature extraction modules (e.g. a Bayesian network for fusing various cues to track people-shaped objects) and internal descriptions (e.g. a log of events relating to people entering or leaving the scene at certain locations and times, along with parameterised models of their visual appearance). Formulating and refining interpretations then amounts to selectively parsing such queries.

3.2 Recognition and classification

Ontologies used in knowledge representation usually consist of hierarchies of concepts to which symbols can refer. Their axiomatisations are either self-referential or point to more abstract symbols. As suggested above, simply defining an ontology for a particular computer vision problem is not sufficient, the notion of how the terms of the ontology are grounded in the actual data is more crucial in practice.

This paper argues 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 [6]. Such criteria can be ranked in a hierarchy and further divided into those which are *necessary* for the object or action to be recognised and those which are merely *contingent*. Such a ranking makes it possible to quickly eliminate highly improbable or irrelevant configurations and narrow down the search window. The combination of individually weak and ambiguous cues 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 information retrieval.

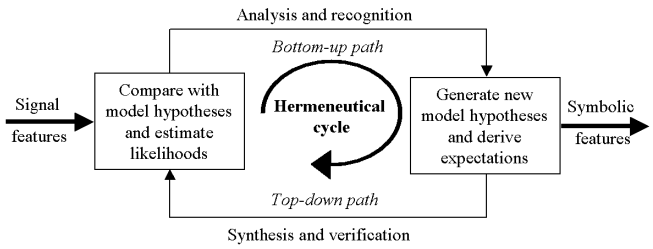

Fig. 1 The Hermeneutical cycle for iterative interpretation in a generative (hypothesise and test) framework.

3.3 Self-referential perceptual inference framework

In spite of the benefits of Bayesian networks and related formalisms, probabilistic graphical models also have limitations in terms of their ability to represent structured data at a more symbolic level [48,47] and the requirement for normalisations to enable probabilistic interpretations of information. Devising a probabilistic model is in itself not enough since one requires a framework that determines which inferences are actually made and how probabilistic outputs are to be interpreted.

Interpreting visual information in a dynamic context is best approached as an iterative process where low-level detections are compared (induction) with high-level models to derive new hypotheses (deduction). These can in turn guide the search for evidence to confirm or reject the hypotheses on the basis of expectations defined over the lower level features. Such a process is well suited to a generative method where new candidate interpretations are tested and refined over time. Figure 1 illustrates this approach.

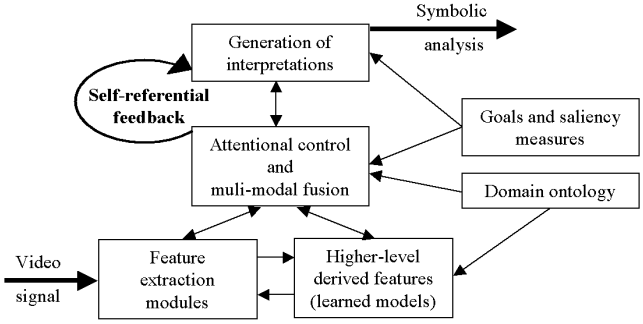

Fig. 2 Sketch of the proposed approach to goal-directed fusion of content extraction modules and inference guided by an attentional control mechanism. The fusion process and selective visual processing are carried out in response to a task and domain definition expressed in terms of an ontological language. Interpretations are generated and refined by deriving queries from the goals and current internal state.

However, there is a need to improve on this methodology when the complexity of the desired analysis increases, particularly as one considers hierarchical and interacting object and behavioural descriptions best defined in terms of a syntax at the symbolic level. The sheer number of possible candidate interpretations and potential derivations soon requires a means of greatly limiting the system's focus of attention. A useful analogy is selective processing in response to queries [8]. Visual search guided by a query posed in a language embodying an ontological representation of a domain allows adaptive processing strategies to be utilised and gives an effective attentional control mechanism.

This paper demonstrates that an ontological content representation and query language could be used as an

effective vehicle for hierarchical representation and goal-directed inference in high-level visual analysis tasks. As sketched in figure 2, such a language would serve as a means of guiding the fusion of multiple sources of visual evidence and refining symbolic interpretations of dynamic scenes in the context of a particular problem. By maintaining representations of both the current internal state and derivational goals expressed in terms of the same language framework, such a system could be seen as performing self-referential feedback based control of the way in which information is processed over time. Visual recognition then amounts to selecting a parsing strategy that determines how elements of the current string set are to be processed further given a stream of lower level tokens generated by feature detectors. The overall structure of the interpretative module is not limited to a particular probabilistic framework and allows context-sensitive parsing strategies to be employed where appropriate.

As shown above, ontologies are gaining popularity for tasks such as multimedia and document annotation. At the same time, many ideas from artificial intelligence and knowledge engineering are being re-formulated using recent advances in probabilistic inference and machine learning, especially as regards the use of Bayesian networks. There has also been recent interest in combining ideas from language processing, information retrieval, and computer vision. This paper builds on many of these ideas and presents a framework for performing visual inference using ontologies. While ontologies often play a passive taxonomic role, the work presented in this paper considers ontologies as an integral part of an active inference framework for computer vision. Furthermore, the ontologies presented here embody both structure (syntax) and meaning (semantics), thus giving rise to the notion of an ontological language. Sentences in such a language are linked to visual evidence by iteratively using the ontology as a structured probabilistic prior to tie together different recognition and processing methodologies. By repeatedly matching and generating ontological sentences, this process becomes increasingly self-referential. The following sections present two computer vision applications that illustrate these concepts.

4 Ontological query language for content-based image retrieval

This section presents a system which allows users to search image databases by posing queries over desired visual content. A novel query and retrieval method called OQUEL (ontological query language) is introduced to facilitate formulation and evaluation of queries consisting of (typically very short) sentences expressed in a language designed for general purpose retrieval of photographic images. The language is based on an extensible ontology which encompasses both high-level and low-level concepts and relations. Query sentences are prescriptions of target image content rather than descriptions. They can represent abstract and arbitrarily complex retrieval requirements at different levels of conceptual granularity and integrate multiple sources of evidence. Further details on OQUEL are available in [65, 61].

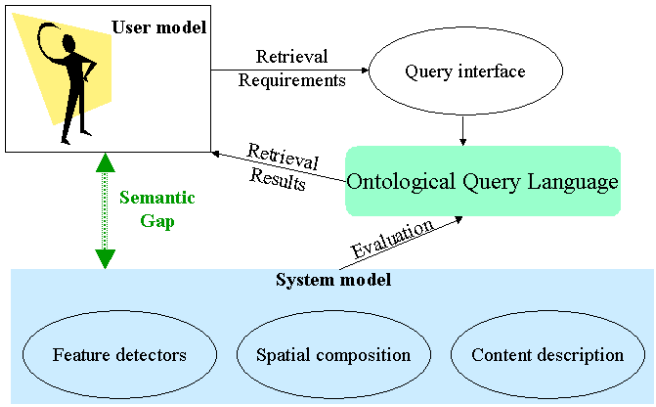

Fig. 3 Model of the retrieval process using an ontological query language to bridge the semantic gap between user and system notions of content and similarity.

The retrieval process takes place entirely within the ontological domain defined by the syntax and semantics of the user query. It utilises automatically extracted image segmentation and classification information, as well as Bayesian networks to infer higher level and composite terms. The OQUEL language provides an effective mechanism of addressing key problems of content based image retrieval, namely the ambiguity of image content and user intention and the semantic gap which exists between user and system notions of relevance (see figure 3). 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. The combination of individually weak and ambiguous cues can be seen as an attempt at extending the success of indicative methods for content representation in the field of text retrieval.

4.1 Syntax and semantics

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. 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". 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:

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

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 (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".

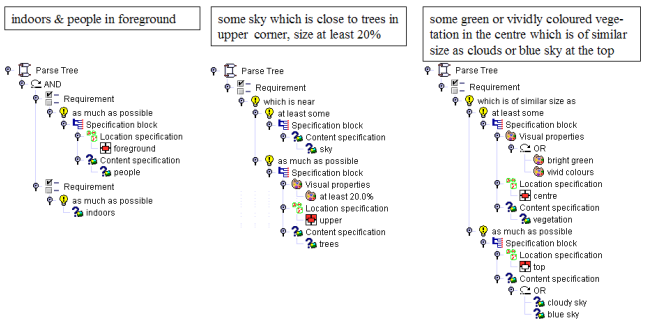

Fig. 4 Examples of OQUEL query sentences and their syntax trees.

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. Figure 4 shows some additional query sentences and their resulting abstract syntax trees.

4.2 Visual content analysis

The OQUEL language has been implemented as part of the *ICON* content-based image retrieval system [63,64]. ICON extracts various types of content descriptors and meta data from images. The following are currently used when evaluating OQUEL text queries:

4.2.1 Image segmentation

Images are segmented into non-overlapping regions and sets of properties for size, colour, shape, and texture

are computed for each region [56,57]. Initially full RGB edge detection is performed followed by non-max suppression and hysteresis edge-following steps akin to the method due to Canny. Voronoi seed points for region growing are generated from the peaks in the distance transform of the initial 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. A texture model based on discrete ridge features is also used to describe regions in terms of texture feature orientation and density. 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.

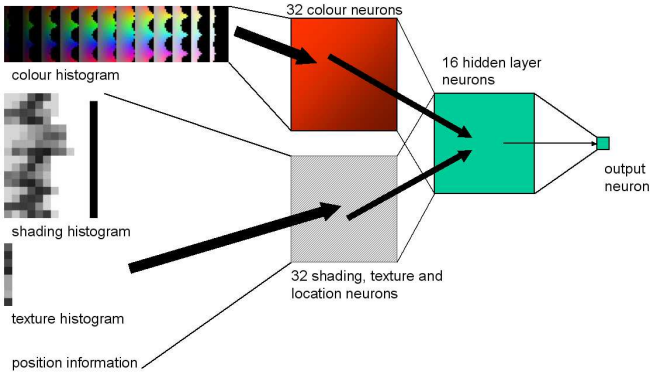

Fig. 5 Example architecture of the neural networks used for image region classification.

4.2.2 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 trained over a large (over 40000 exemplars) corpus of manually labelled image regions. Figure 5 shows an example of the MLP network structure. Evaluation results from the test set were used to obtain the classifier confusion matrix shown in table 1. The numbers along the main diagonal represent the probabilities of correct classification $P(c_i|c_i)$ while the other entries give the probability $P(c_j|c_i); i \neq j$ of a region of class c_i being erroneously classified as belonging to class c_j.

Automatic labelling of segmented image regions with semantic visual categories [63] such as grass or water that 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 that can be used to ground higher level concepts and content prescriptions.

4.2.3 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 [37], 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.

4.2.4 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 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. A nearest-neighbour shape classifier was trained to recognise eyes. 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.

4.2.5 Content representation

After performing the image segmentation and other analysis stages as outlined above, image content is represented at the following levels:

$P(c_j\|c_i)$												
c_i		c_j										
i	c_i label	c_0	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9	c_{10}	c_{11}
0	Skin	**0.78**	0	0.01	0	0	0	0	0	0.12	0	0.09	0
1	Blue sky	0	**0.80**	0.12	0	0	0	0	0	0	0	0	0.08
2	Cloudy sky	0	0	**0.75**	0	0	0.04	0	0.05	0	0	0.12	0.04
3	Snow	0	0.07	0.06	**0.87**	0	0	0	0	0	0	0	0
4	Trees	0	0	0	0	**0.83**	0.14	0	0.01	0	0.02	0	0
5	Grass	0	0.03	0.01	0	0.22	**0.73**	0	0.01	0	0	0	0
6	Tarmac	0.04	0	0.02	0	0.02	0	**0.59**	0.11	0	0.04	0.12	0.06
7	Water	0	0.03	0.05	0.08	0.01	0.06	0.01	**0.64**	0	0.02	0.06	0.04
8	Wood	0.02	0.01	0	0	0	0	0.02	0	**0.71**	0.02	0.22	0
9	Brick	0.02	0	0	0	0.05	0	0.02	0	0.04	**0.79**	0.08	0
10	Cloth	0	0	0	0	0	0	0	0.10	0.07	0.03	**0.76**	0.04
11	Int.Walls	0.04	0	0.04	0	0	0	0	0	0.02	0.08	0	**0.82**

Table 1 Region classifier confusion matrix $\mathbf{C_{ij}} = P(c_j|c_i)$.

- *Region mask*: Canonical representation of the segmented image giving the absolute location of each region by mapping pixel locations onto region identifiers.
- *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.
- *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 grid). For each grid element there consequently is 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.

Through the relationship graph representation, matching of clusters of regions is made invariant with respect to displacement and rotation using standard matching algorithms. The grid pyramid and region mask representations allow an efficient comparison of absolute position and size.

4.3 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. This section discusses 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 above. Other terminals (modifiers, scene descriptors, binary relations, and semantic categories) and syntactic constructs are evaluated by the query parser as will be discussed in section 4.4.

- *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 regions 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 the 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").

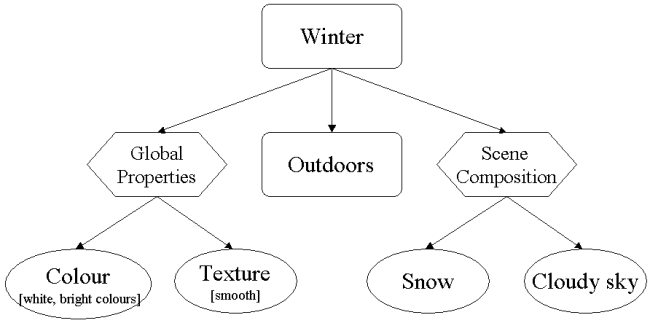

Fig. 6 Simplified Bayesian network for the scene descriptor "winter".

4.4 Query evaluation and retrieval

This section discusses the OQUEL retrieval process as implemented in the ICON system. OQUEL queries are parsed to yield a canonical abstract syntax tree (AST) representation of their syntactic structure. Figures 4, 7, 8, 9, and 10 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 shape descriptors 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.

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 one arrives at a single probability of relevance for the entire image. At the lowest level, tokens map directly or very simply onto the content descriptors via SQL queries. 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 an estimate of whether people are present in the image.

Matching image content is retrieved and the initial list of results is sorted in descending order of a probability of relevance score. Next, nodes denoting visual properties (e.g. size or colour) are assessed in order to filter the initial results and modify relevance scores according to the location, content, and property specifications which occur in the syntax tree. Finally, relationships (logical, geometric, or semantic, e.g. similarity) are assessed and probability scores are propagated up the AST until each potentially relevant image has one associated relevance score. Relations are evaluated by considering matching candidate image content (evidence) . 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. Query sentences consist of requirements which yield matching probabilities that are further modified and combined according to the top level syntax.

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 additional 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. Figure 6 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.

Due to the inherent uncertainty and complexity of the task, query evaluation is performed in a way that limits the requirement for runtime inference by quickly ruling out irrelevant images given the query. The overall approach 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 [45]. The approach adopted here is related to that of [33] in that it applies notions of weighting akin to the Dempster-Shafer theory of evidence to construct an information retrieval model which captures structure, significance, uncertainty, and partiality in the evaluation process. 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

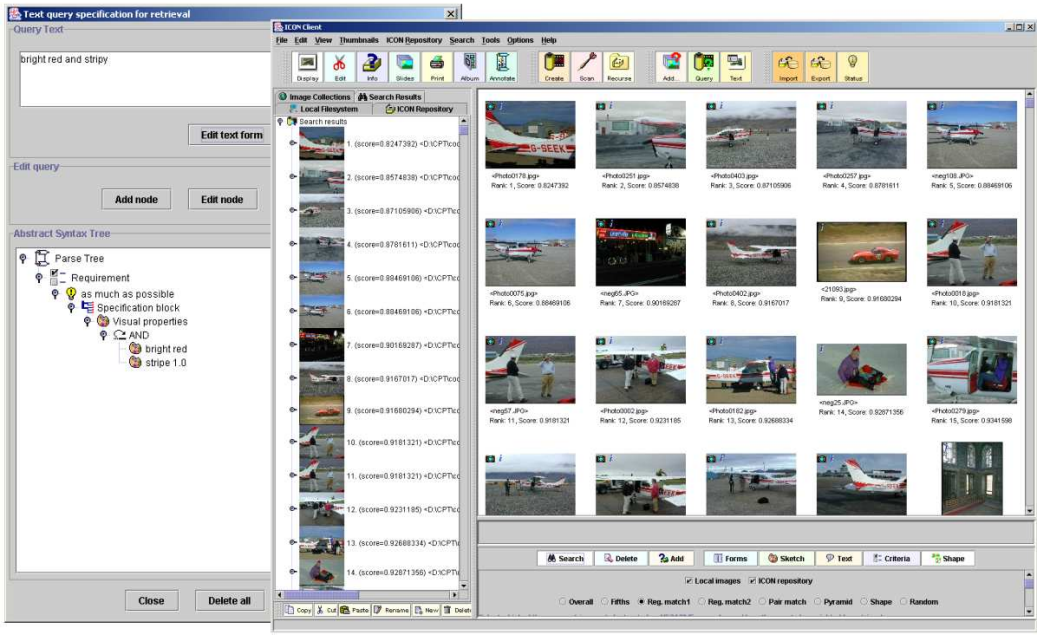

Fig. 7 Search results for OQUEL query A "bright red and stripy".

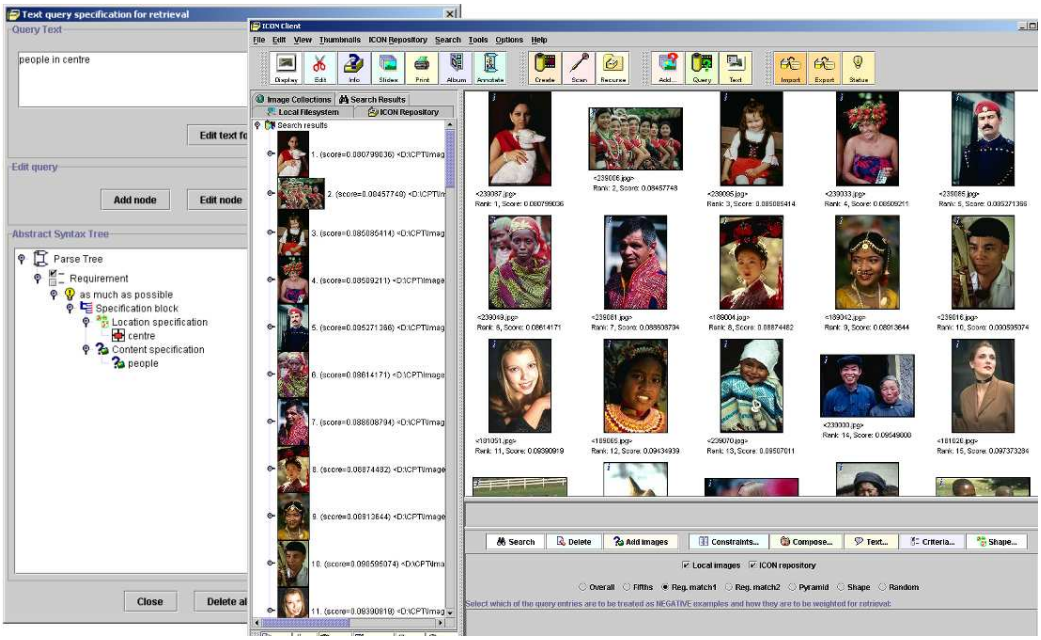

Fig. 8 Search results for OQUEL query B "people in centre".

[18]. Fuzzy logic offers a principled way of mapping linguistic terms to probabilities via fuzzy sets and has been successfully applied to CBIR in the past ([18,41,34]). 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.

Relevant images are those where one 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 the system

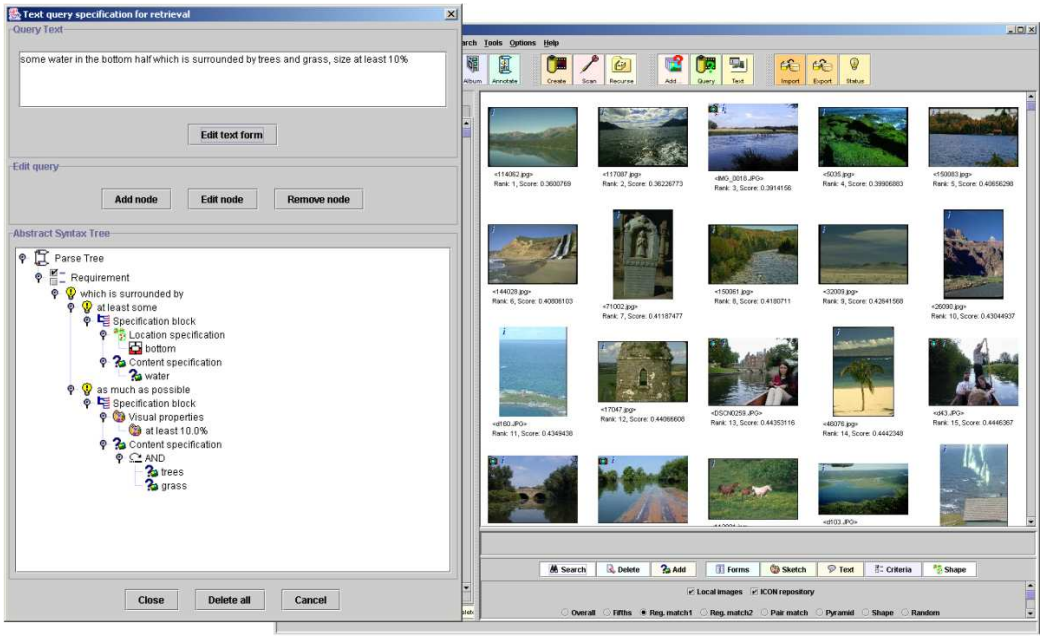

Fig. 9 Search results for OQUEL query C "some water in the bottom half which is surrounded by trees and grass, size at least 10%".

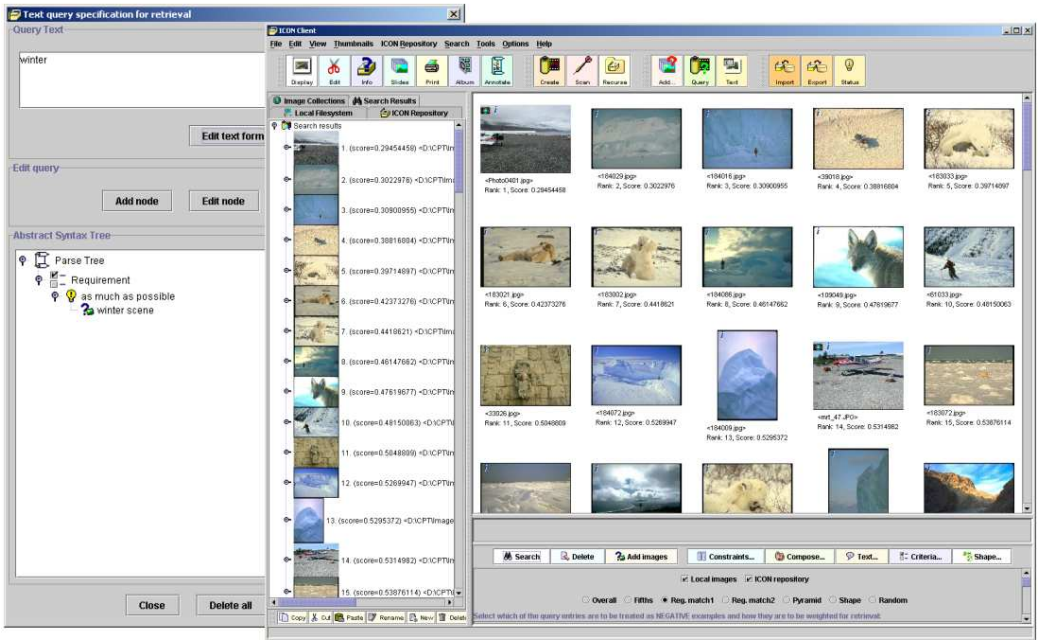

Fig. 10 Search results for OQUEL query D "winter".

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 one maintains a notion of uncertainty, borderline false detections will simply result in lowly ranked retrieved images. Top query results correspond to those image where the confidence of having found evidence for the presence of people is high relative to the other images.

4.5 Qualitative and quantitative evaluation

While most evaluation of CBIR systems is performed on commercial image collections such as the Corel image

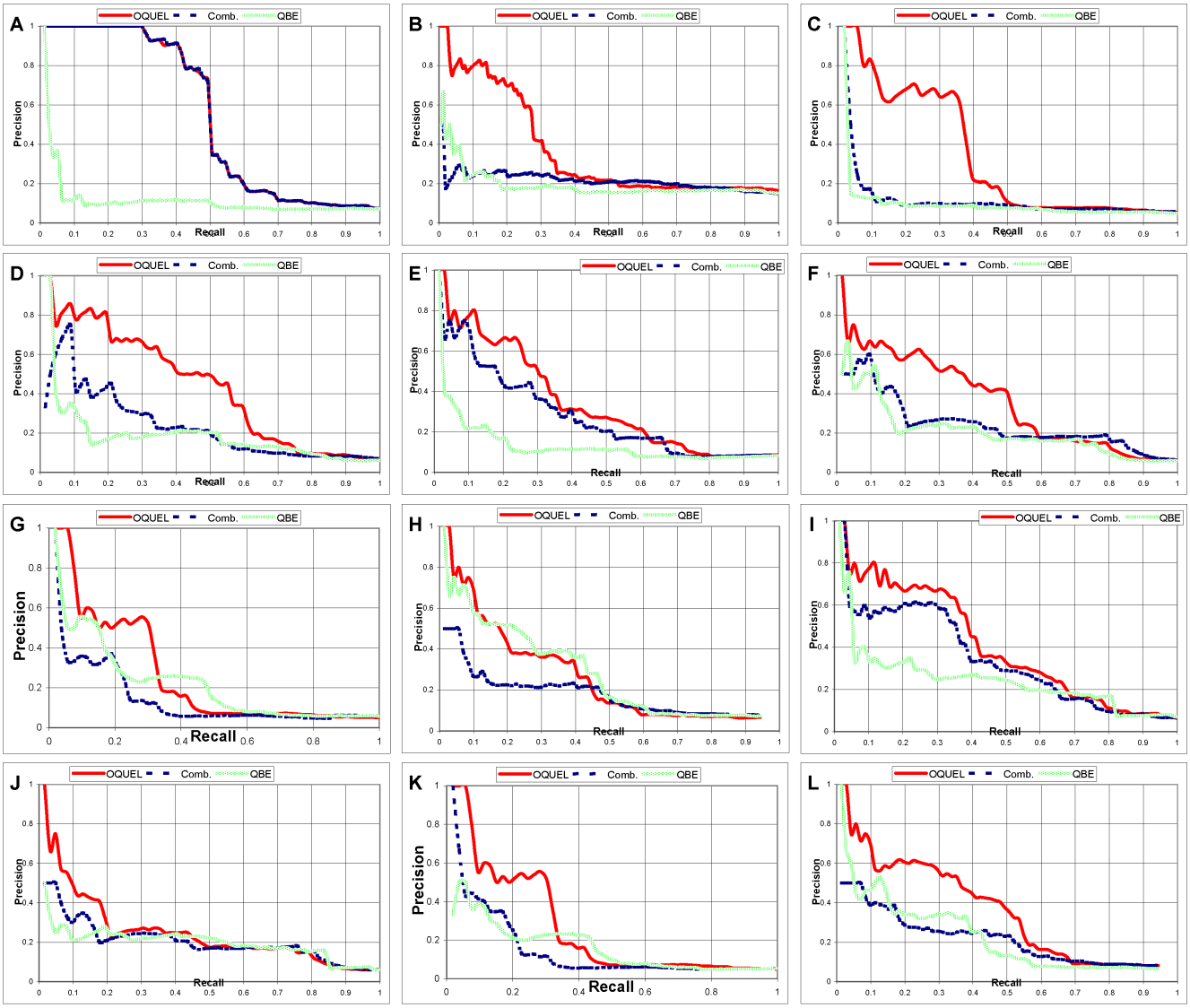

Fig. 11 Plots of relative percentages for precision versus recall for the retrieval experiments. As can be seen, results obtained using OQUEL (red) generally outperform those achieved using query-by-example (green) or a combination of sketch and feature based retrieval (blue).

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 [38]. Therefore a set of images consisting of 670 Corel images augmented with 412 amateur digital pictures of highly variable quality and content were chosen. Manual relevance assessments in terms of relevant vs non-relevant were carried out for all 1082 images over the test queries described below. Twelve retrieval requirements were chosen, 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"
Query E "artificial stuff, vivid colours and straight lines"
Query F "indoors & people in foreground"
Query G "some vividly coloured vegetation in the centre which is of similar size as clouds or blue sky at the top"
Query H "city or countryside"
Query I "artificial stuff and tarmac"
Query J "people or animals"
Query K "[blue sky at top] and [trees centre] and grass lower"
Query L "some sky which is close to buildings above"

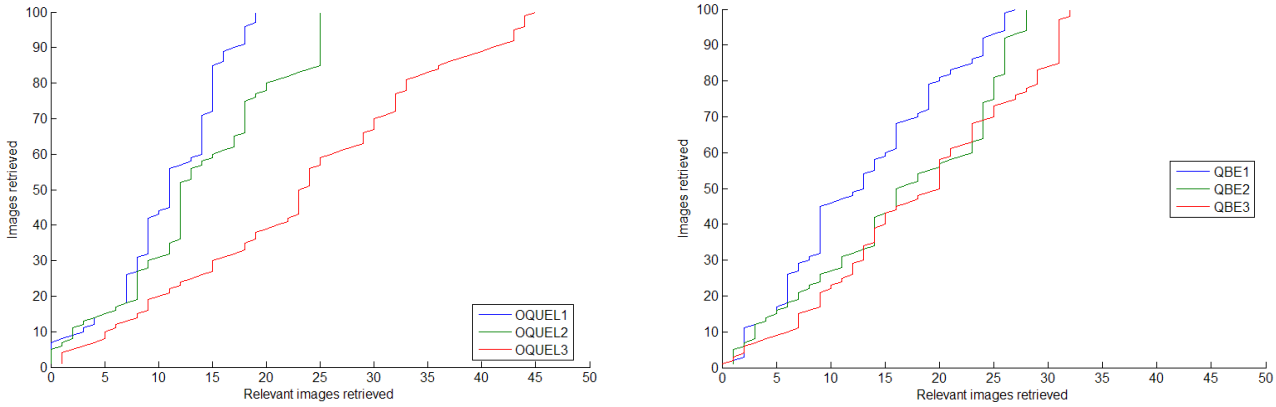

Fig. 12 Plots of total number of images retrieved versus number of relevant images retrieved for *left*: OQUEL queries, *right*: query-by-example (QBE). In each case, results are shown for an initial query and two iterations of query refinement.

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. For the first four OQUEL queries, top ranked search results are shown in figures 7, 8, 9, and 10. For each OQUEL query a further two queries embodying the same retrieval need were composed using the other search facilities of the ICON system:

- Combined query ("Comb."): a query which may combine a sketch with feature constraints as appropriate to yield best performance in reasonable time.
- Query-by-example ("QBE"): 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.

To quantify performance, graphs of precision versus recall were computed using manual relevance assessments for each test query as shown in figure 11. In each case, OQUEL query results are shown together with results for the two other query modalities described above, i.e. a combined query ("Comb.") and a query-by-example ("QBE") designed and optimised to meet the same user search requirements. It can be seen that OQUEL queries generally 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 invariances, or object level constraints.

As recommended in [39], the *normalised average rank* was also computed (see table 2) which is a useful stable measure of relative performance in CBIR:

	$Rank^{\sim}$		
Query	OQUEL	Comb	QBE
A	0.2176	0.2175	0.3983
B	0.2915	0.3072	0.3684
C	0.2628	0.3149	0.3521
D	0.1935	0.2573	0.2577
E	0.2152	0.2418	0.3324
F	0.1969	0.1816	0.2475
G	0.3147	0.3766	0.2831
H	0.3312	0.2947	0.2952
I	0.1863	0.2123	0.2105
J	0.2170	0.2113	0.2020
K	0.3151	0.4078	0.3377
L	0.2367	0.2558	0.3351

Table 2 Results of the query experiments indicating the normalised average rank measure for each of twelve query experiments (A..L) and for each of three methods of query composition (OQUEL, "combined", and "query-by-example").

$$Rank^{\sim} = \frac{1}{N N_{rel}} \left[\sum_{i=1}^{N_{rel}} R_i - \frac{N_{rel}\,(N_{rel}+1)}{2} \right] \qquad (1)$$

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

4.6 Scalability and query refinement

In order to investigate the scalability of the OQUEL retrieval technology, an image collection consisting of over 12000 high-resolution photographic images was compiled.

The images were taken by 11 different amateur photographers and represent a very diverse range of subject matter, focal lengths, lighting conditions, and picture quality. Many of the images were taken indoors, are poorly lit, or blurred. Several of them are upside down or rotated by 90^o, which can cause additional problems for CBIR systems which rely on spatial composition.

An important goal of CBIR is to allow users to identify sets of images which are semantically related yet disparate in their visual properties and composition. In order to test the suitability of the OQUEL language for such a task, a retrieval requirement for images of archaeological sites was chosen as a test case. The collection of 12000 images does indeed contain several images which meet this broad description, taken at diverse locations across the globe and featuring a variety of different styles, periods, and surroundings (e.g. ancient buildings in a modern city, ruins in the desert or jungle). The retrieval requirement was translated into an initial OQUEL query which was subsequently modified twice in light of search results. This allows the ease and effectiveness of query refinement within the OQUEL framework to be assessed. In order to avoid the prohibitive effort of manually assessing and ranking every image in the collection, only the top 100 images returned by each query were analysed and rated as being either relevant or not relevant with respect to the task of finding pictures of archaeological sites. Most users are unlikely to view more than the top 100 results [49] and this method is sufficient for quantitative comparison of the relative merits of different approaches. The following OQUEL queries were searched on using the ICON system:

- *OQUEL1* (*initial query*): "brick and (grass or trees)"
- *OQUEL2* (*first refinement*): "[outdoors] and brick"
- *OQUEL3* (*second refinement*): "[outdoors] and [summer] and brick"

Note that the OQUEL language does not currently feature semantic terms characterising buildings and hence the query had to be re-expressed in simpler terms. In order to quantify precision by means of the cumulative frequency of relevant images returned by each query, figure 12 shows results in terms of the number of images retrieved versus number of relevant images retrieved for the top 100 search results. It can be seen that even simple refinement of the OQUEL queries leads to improvements in performance without requiring complicated queries. In order to contrast the performance of OQUEL on this task with another retrieval method, queries were also composed by selecting example images. Results for these are also shown in figure 12. After some manual browsing, a relevant image was found and used as a single positive example forming the first query (QBE1). Subsequently one non-relevant image was selected from the QBE1 retrieval results and added to the query to form a new query (QBE2). Finally, an additional relevant image was added to the query set to form QBE3. As can be seen, absolute performance is significantly lower and even the refined QBE queries fail to adequately capture the semantics behind the retrieval requirement, even though all queries have access to the same set of image descriptors.

4.7 Conclusions

Comparisons with other query composition and retrieval paradigms implemented in ICON (sketch, sample images, property thresholds) show that the OQUEL query language constitutes a more efficient and flexible retrieval tool (see table 2). 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. Such queries have also generally proven to be more efficient to evaluate since one only needs to analyse those aspects of the image content representation that 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.

Future work will seek to extend the ontology and increase the scope for quantitative analysis by adapting the OQUEL framework to work with emerging CBIR evaluation efforts. For example, TRECVID[1] has specified a number of data collections and search tasks suitable for CBIR evaluation. It features a collection of several thousand video keyframes and a set of search tasks and relevance judgments for evaluation. Most target queries are formulated as user requirements (e.g. "I need some clips showing Glen Canyon dam") and therefore need to be translated into the chosen query modality. At the moment, queries generally require one to search for a specific object or person (e.g. "David. J. Nash."), category (e.g. "Football players"), or activity (e.g. "leisure time at the beach"). Many such queries therefore require very specific factual knowledge that is best obtained through associated data such as closed-captions or speech transcription (which is provided by recent TRECVIDs). At the same time, OQUEL provides relational constructs (e.g. "larger than", "similar colour") that are not generally required by TRECVID queries at present. Ultimately, there is no such thing as a typical CBIR query, but OQUEL is able to span the gamut between factual keyword-based queries and more complex queries with a rich syntax. It is likely that query languages such as OQUEL will prove useful as image and video retrieval methods merge with text-based information retrieval in order to keep up with ever increasing user demand for easy access to today's vast document and multimedia collections.

[1] *http://www-nlpir.nist.gov/projects/trecvid/*

5 Ontology-guided dynamic scene understanding

This chapter presents work showing how the process of creating recognition systems for high-level analysis of surveillance data can be largely automated, provided sufficient quantities of training data (ground truth) which has been annotated with descriptors from the desired analysis specification are available. Such a specification may usefully be regarded as an ontology which provides a prior description of the application domain in terms of those entities, states, events and relationships which are deemed to be of interest. The hierarchical organisation and relational constraints imposed by the ontology can then be used to guide the design of a complete visual analysis system.

In this chapter, video sequences and ground truth from the CAVIAR project[2] were used to define an ontology of visual content descriptors arranged in a hierarchy of scenarios, situations, roles, states, and visual properties. The latter properties were defined by choosing object attributes such as translational speed and appearance change which could easily be computed by means of a visual tracking and appearance modelling framework. The CAVIAR training data was then automatically re-labelled with this extended set of descriptors by instantiating the tracking framework with the individual objects in the ground truth and computing the selected visual attributes for all frames in the sequences. The resulting data was then used to learn both the structure and parameters of Bayesian networks for high-level analysis. Evaluations were performed to assess how easily the categories of the ontology could be inferred on the basis of the chosen visual features and on the basis of preceding layers in the hierarchy. The former allows one to assess the (in)adequacies of a set of given visual content extraction and representation methods, which is an important tool in designing the computer vision components of a surveillance system in order to maximise their utility for high-level inference in light of the domain ontology. Conversely, one can use the probabilistic scoring methods applicable to Bayesian networks to evaluate how well-defined e.g. the pre-defined set of situation descriptors are in terms of the labels for object roles and states which appear in the ground truth. Further implementation details can be found in [62,61].

5.1 Visual analysis and tracking

The tracking system maintains a background model and foreground motion history (obtained by frame differencing) which are adapted over time using an exponential rate of decay to determine the decreasing influence of previous frames im_{i-1} in the history. The motion history M_i is used to identify a background image bim_i of pixels undergoing sufficiently slow change which can then be used to reliably update the background model B_i and estimate its variance:

$$B_i = \alpha * bim_i + (1-\alpha) * B_{i-1}; \quad B_0 = im_0 \tag{2}$$

$$where\ bim_i = |im_i - M_i| < \tau; \quad \alpha = 1 - e^{-1/\lambda_B} \tag{3}$$

Pixels are deemed to be part of the dynamic foreground if they exceed a difference threshold which is a multiple of the background variance σ_i^B and if they are not deemed to be part of a shadow as determined by a simple test over luminance vs hue and saturation changes.

After performing some morphological clean-up operations, foreground pixels are clustered using connected components analysis to identify moving regions ("blobs"). Blob positions are tracked using a Kalman or particle filter with a second order motion model. Tracked objects are matched to detected blobs using a weighted dissimilarity metric which takes into account differences in predicted object location vs blob location and changes in object appearance. Colour appearance is modelled by histograms in RGB space and Gaussian Mixture models in HSV space, with distance metric computed through the Earth Mover's distance and likelihood computation respectively. Object arrivals, departures and occlusions are inferred using a Bayesian network (see figure 14). Figure 13 illustrates the approach.

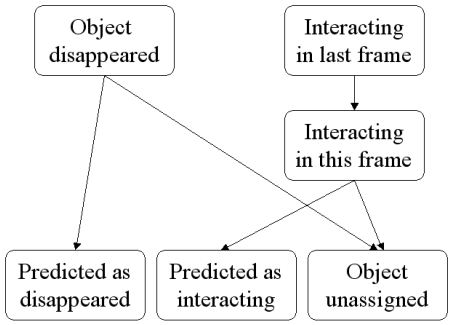

Fig. 14 Bayesian network for occlusion reasoning and prediction of object interactions.

In order to parameterise object motions and deformations, we have adapted a sample-based edge tracking method due to Smith [58]. The method samples points along edges, tracks them over subsequent frames, and estimates a parameterised motion model for the whole edge.

In order to model and track the shape of objects in terms of their closed boundaries, approaches based on active contours have been very prominent in the vision community. We have adapted the Gradient Vector flow (GVF) method of Xu and Prince [70] in order to make the transition from tracked edges to closed boundary contours. The snake's external force is computed as a diffusion of the gradient vectors of an edge map derived using our edge detector.

[2] EC Funded CAVIAR project/IST 2001 37540, *http://homepages.inf.ed.ac.uk/rbf/CAVIAR/*.

Fig. 13 Tracking results (from left to right): Original frame; Model of the background variances; Results of background subtraction; Detected blobs after morphological operations; Resulting tracked objects (outlined in green) with ground truth data and results shown in yellow.

After fitting a motion model to the edges of each object by means of the edge tracker and also fitting the GVF active contour to the object's outer boundary, the shape of that boundary is parameterised. The approach chosen here uses the first four of the seven affine invariant moments ϕ_i proposed by Hu [24]. In the present case, the invariant moments are calculated over object boundary pixels only, i.e. $f(x, y)$ is an indicator function which is 1 for a given object's boundary pixels and 0 elsewhere. All x, y values are transformed such that they lie in the range $[0; 1]$ by normalising with respect to the object's bounding box. In order to reduce the range of the ϕ_i, the logarithms $log(\phi_i)$ of the actual values are used.

Scenario	A description of an individual's overall context.
scBSC	Browsing scenario
scIM	Immobile scenario
scWG	Walking scenario
scDD	Drop-down scenario
Situation	The situation in which the individual is participating.
siM	Moving situation
siIS	Inactive situation
siBSI	Browsing situation
Role	The individual's role in the current situation.
rF	Fighter role
rBR	Browser role
rLV	Left victim role
rLG	Leaving group role
rWR	Walker role
rLO	Left object role
State	The individual's current attributes.
tAP	Appear
tDI	Disappear
tO	Occluded
tIN	Inactive: visible but not moving
tAC	Active: visible, moving but not translating across the image
tWK	Walking: visible, moving, translating across the image slowly
tR	Running: visible, moving, translating across the image quickly

Fig. 15 CAVIAR ontology

5.2 Ground truth and domain ontology

Video sequences and ground truth from the CAVIAR project [3] comprise 28 annotated sequences taken by a surveillance camera in the entrance lobby of the INRIA Rhone-Alpes research laboratory in France. They consist of six scenarios of actors performing different activities such as walking around, browsing information displays, sitting down, meeting one another and splitting apart, abandoning objects, fighting and running away. The CAVIAR annotations can naturally be organised into a hierarchical ontology as shown in figure 15. This arrangement offers guidance for the design of Bayesian inference networks. For example, one would expect an individual's state to depend primarily on their current role, their current role to depend on the situation they are facing, and their situation to depend on the scenario in which they are participating. These relationships can be used as a structural prior for the training of Bayesian networks as described below.

In order to ground the terms of the ontology, we extend it with appropriate descriptors computed from the tracking and appearance modelling framework described in section 5.1. These visual descriptors (see table 3) are not claimed to constitute the best choice for the analysis task at hand. They are merely properties of tracked objects which can be simply and robustly defined using the techniques described in section 5.1 and offer a reasonable basis for studying the requirements for low-level analysis mechanisms which result from the pre-defined ontology of higher-level terms.

5.3 Learning Bayesian network structure and parameters

There are a variety of methods for learning both the parameters and structure of Bayesian networks from data, see e.g. [28]. In this paper, the goal was to learn the structure and parameters of a static directed Bayesian network given fully observed data, i.e. the values of all nodes are known in each case from the ground truth (augmented as required with the information gathered

[3] EC Funded CAVIAR project/IST 2001 37540, see http://homepages.inf.ed.ac.uk/rbf/CAVIAR/

Name	Explanation
CVx	Relative x-position
CVy	Relative y-position
CVv	Speed
CVa	Absolute acceleration
CVm	Relative mass
CVn	Relative change in mass
CVt	Major axis orientation
CVrx	Direction of movement relative to positive x-axis
CVry	Direction of movement relative to positive y-axis
CVf	Motion flow (exponentially smoothed history of the object's motion)
CVl	Object lifetime
CVs	Combined appearance measure difference score
CVo	Occlusion status
CVbm	Six element (CVbm1..CVbm6) vector of motion model parameters corresponding to the projective deformations of x- and y-translation, rotation, dilation, pure shear, and shear at 45^o
CVbs	Four element (CVbs1..CVbs4) vector of shape model parameters corresponding to the invariant moments $(\phi_1, \phi_2, \phi_3, \phi_4)$

Table 3 Extended set of computer vision derived nodes

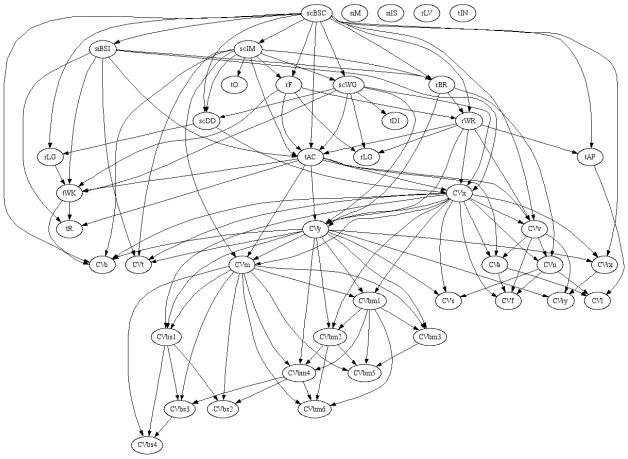

Fig. 16 Bayesian network structure trained using the K2 algorithm from the extended ontology.

by the computer vision techniques). The variables in table 3 were discretised by choosing a number of quantisation levels (usually 3 or 4) and quantising by sub-dividing the range $[\mu - 2\sigma; \mu + 2\sigma]$ (where μ and σ are the mean and standard deviation respectively of the observed values of the variable as computed over the entire data set) into the corresponding number of subranges. Each value of the variable in the data set is then quantised by assigning it one of the quantisation values according to the subrange which it occupies. Making the reasonable assumption that the values are approximately normally distributed, this quantisation method accounts for about 95.5% of the variation in the data while reducing the

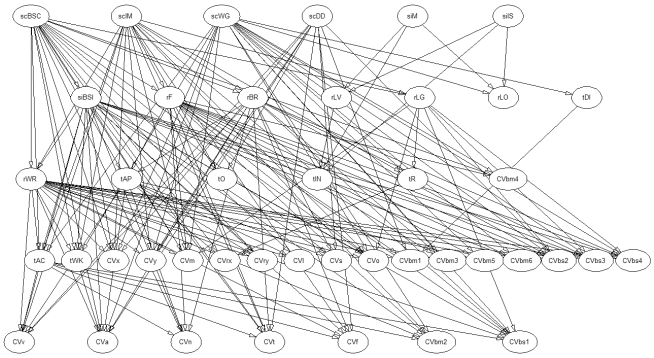

Fig. 17 Bayesian network structure trained using the K2 algorithm with a structural prior from the extended ontology.

effect of outliers that may occur due to discontinuities caused by imperfect visual analysis.

Learning the network structure requires a means of searching the space of all possible DAGs over the set of nodes X and a scoring function to evaluate a given structure over the training data D. Two different learning algorithms were chosen and implemented by means of the Bayes Net Toolbox for Matlab [40].

The K2 algorithm [11] is a greedy search technique which starts from an empty network but with an initial ordering of the nodes. A Bayesian network is then created iteratively by adding a directed arc to a given node from that parent node whose addition most increases the score of the resulting graph structure. This process terminates as soon as none of the possible additions result in an increased score.

Markov Chain Monte Carlo (MCMC) is a family of stochastic search methods. As described in [17], MCMC can be applied to Bayesian network structure learning without the need for a prior node ordering (although such orderings can be employed to speed up convergence). The Metropolis-Hastings sampling technique is applied to search the space of all graphs G by defining a Markov Chain over it whose stationary distribution is the posterior probability distribution $P(G|D)$. Following Bayes' rule, $P(G|D) = P(D|G)P(G)$. The marginal likelihood of the data $P(D|G)$ can be computed by means of an appropriate scoring function (see below) and the prior $P(G)$ may be left uninformative (i.e. a uniform distribution over the set of possible DAGs G). Candidate structures are then sampled by performing a random walk over the Markov chain. The highest scoring network structure can then be inferred by averaging over a sufficiently large number of samples.

In order to compute the score of a candidate network over the training data while avoiding overfitting, two scoring functions were considered:

– The marginal likelihood of the model

$$P(D|G) = \int_{\theta} P(D|G,\theta)P(\theta|G)$$

where D is the training data, G is the graph structure, and θ are the network parameters.

– The Bayesian Information Criterion (BIC), which approximates the marginal likelihood using a Minimum Description Length (MDL) approach. Following [21], the Laplace approximation to the parameter posterior can be written in terms of the likelihood and a penalty term $\frac{d}{2}\log M$ to explicitly penalise model complexity:

$$\log P(D|G) \approx \log P(D|G,\hat{\theta}_G) - \frac{d}{2}\log M = \mathrm{BIC}(D,G)$$

where M is the number of training cases in D, d is the number of free parameters, and $\hat{\theta}_G$ is their maximum likelihood estimate.

MCMC was largely found to provide inferior results and required many thousands of iterations to converge to a solution. Furthermore, the K2 method directly benefits from the prior structural information contained in the ontology. Although the BIC score is a more crude approximation than that inherent in the computation of the marginal likelihood shown above, there was very little difference in resulting network performance using the two scoring methods. Once the network structure has been trained, parameters can be estimated easily using maximum likelihood estimation using Dirichlet priors (pseudo-counts).

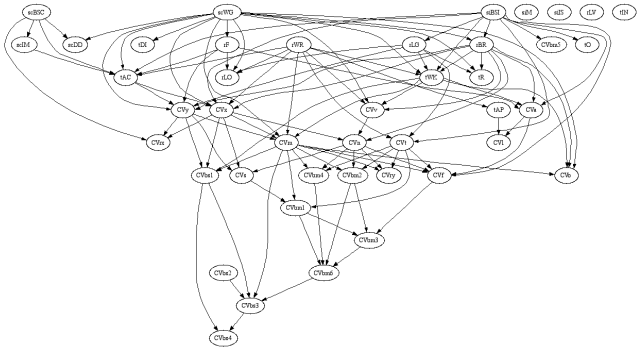

Fig. 20 Bayesian network structure trained using the K2 algorithm with random initial ordering of the nodes.

5.4 Results

Figure 16 shows a Bayesian network trained in the manner described in section 5.3 using the K2 algorithm without a structural prior. The network structure looks somewhat erratic but captures some of the hierarchical relationships between variables that one would expect from their semantics. Some nodes in the network (siM, siIS, rLV, tIN) remain unconnected. That is because their values are almost constant in the data set and hence can in most cases be inferred trivially through a purely deterministic prior.

Fig. 21 Bayesian network structure trained using the MCMC algorithm initialised with a structural prior.

The network shown in figure 17 was trained using K2 and a structural prior specifying that nodes which are part of the same semantic level in the ontology (e.g. all situation labels) should be treated as equivalent in terms of the ordering of nodes. The resulting network structure encompasses many of the causal relationships one would expect from the semantics and shows that there are strong dependencies between the computer vision derived terms and the states and roles in particular. Figure 18 shows classification rates achieved by this network given different sets of evidence. As shown in figure 19, this network achieves better recognition rates than the one trained without a structural prior. The average recognition rates (computed over an independent testing set) over all elements of the CAVIAR ontology are now 0.911 and 0.882 respectively for the Bayesian network with and without a structural prior.

Even without the structural prior (which groups variables in the same level of the ontology into an equivalence class), the ontology has thus far still been used to define an ordering of the nodes in the Bayesian network. Figure 20 shows an example of a Bayesian network which was trained using K2 without any such ontological information, i.e. with a random initial ordering of the nodes. The performance of this network is worse than the two discussed before, with the average recognition rate now

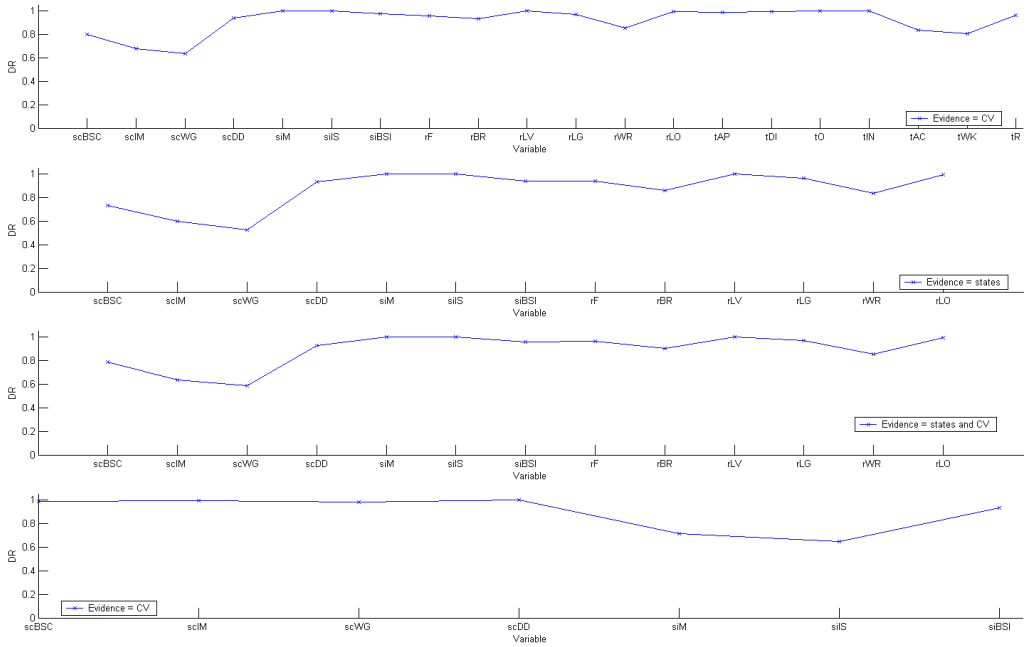

Fig. 18 Plot of recognition rates achieved by the Bayesian network shown in figure 17 for the variables in the CAVIAR ontology given different evidence. *Top*: **given only the computer vision derived nodes as evidence;** *2nd from top*: **given only the states (tAP, tDI, tO, tIN, tAC, tWK, tR);** *3rd from top*: **given both the states and computer vision information;** *Bottom*: **recognition rates for the states given the computer vision derived nodes as evidence. It can be seen that some nodes in the ontology are insufficiently grounded using the computer vision components alone but that embedding them in an ontology derived structure improves recognition.**

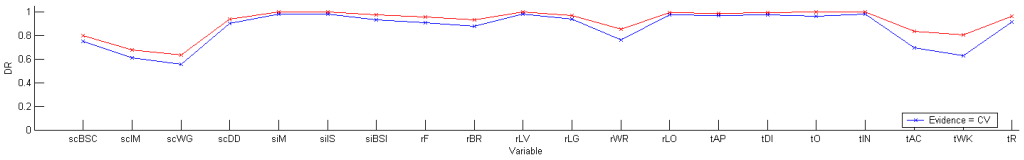

Fig. 19 Plot of Bayesian network recognition rates for the variables in the CAVIAR ontology given only the computer vision derived nodes as evidence. The rates achieved by the network in figure 17 are shown in red (top line), those for the network in figure 16 are shown in blue (bottom line). It can be seen that the use of an ontological prior improves accuracy.

being 0.865. This implies that the use of an ontological prior has reduced the expected error rate from about 14% to about 9% (i.e. a 36% reduction) even though the same data set, image processing, and learning method are being used.

It is also possible to provide a structural prior for MCMC learning by initialising the search with a particular network structure. Figure 21 shows a network trained using a MCMC process whose starting point was the network structure shown in figure 17. The resulting average recognition rates for the CAVIAR ontology are 0.642 (MCMC without ontological prior) and 0.659 (search initialised with a prior, figure 21).

6 Conclusions

6.1 Summary

This paper presents research in the area of high-level computer vision which shows how ontologies can be used as effective computational and representational mechanisms that allow one to relate semantic descriptors to their parametric representations in terms of the underlying data primitives. A particular focus of this work is on the role of ontologies as a means of representing structured prior information and of fusing different kinds of information in an inference framework.

Section 4 presents a novel approach to content-based image retrieval founded on an ontological query language, OQUEL. The problems of expressing, representing, and matching user queries are thus solved through a *prescrip-*

tive ontology of image content descriptors which is hierarchically decomposed using a language which embodies a general syntax and semantics for query composition and representation of target image content. Unlike most conventional "query-by-example" or "query-by-sketch" retrieval interfaces, OQUEL does not require users to select or generate a concrete instantiation of the desired image content and concepts. The language is concise and abstract without being inflexible or overly formal. Query sentences are grounded through a range of image analysis methods that represent image content at low, intermediate, and high semantic levels. This is realised using segmented region properties, classifiers built upon the region parameterisation, and Bayesian inference networks respectively.

It is shown how the ontological query language provides a way of narrowing the *semantic gap* between users and the retrieval system by providing a shared language and hierarchy of concepts for both. Rather than attempting to describe image content in terms of the language, this approach recognises that the meaning attributed to a given image by a user relative to some current retrieval need (and therefore its relevance to a given query) is only discernable through the composition of the query itself which defines the ontological domain over which relevance assessment is carried out. Inference of image content thus occurs only directly in response to the user query and terminates as soon as the relevance or irrelevance of each image has been established. The central role of the ontology is to provide a means for users to define the ontological domain of discourse and for the system to execute the query by grounding and assessing the particular ontological sentence with respect to the actual image data. The syntactic and semantic relationships and redundancies of the ontology, the OQUEL queries, and of image content provide a basis of inference and contextual disambiguation through which the ontological language can be extended with new terms.

In section 5, the problem of building reliable high-level recognition systems for dynamic scene analysis (in particular that of surveillance video) is addressed by a combination of pre-annotated training data, a set of automatically derived visual descriptors, and an extended ontology incorporating both of these. The section describes how Bayesian networks can be trained from this data to perform inference over the terms of the ontology. Moreover, an analysis of the composition and performance of different Bayesian recognition networks can lead to insights into the coherence, utility, and groundedness of the ontology itself in terms of the basis vocabulary derived by the visual analysis.

Ontology in this case is used in a *descriptive* capacity with grounding of the higher level descriptors occurring through statistical learning from the annotated examples and additional features derived by a range of computer vision tracking and visual analysis methods. The hierarchical organisation of the ontology directly adds value to the process by serving as a structural prior which improves the performance of the Bayesian networks. As in section 4, knowledge about the domain is encoded both *intensionally* through the syntactic relationships between terms of the ontology, and *extensionally* by means of the visual processing modules and Bayesian inference networks that were trained to recognise these terms from annotated ground truth.

6.2 Discussion

A central problem in the development and application of ontologies is that of grounding their terms and relations in the underlying data. One way in which this may be achieved is to hierarchically decompose and re-express the terms of the ontology until they are all defined in terms of primitives which the system can readily recognise. Another way is to provide sufficient training data such that the system can be made to internalise an appropriate definition of the concept by means of machine learning. Both of these approaches are investigated in this work.

Furthermore, the notion of ontology based languages is introduced as a powerful means of creating computational vehicles for knowledge representation and matching which incorporate the syntactic and semantic structures characterising a given domain. A further approach put forward is the concept of visual analysis as a dynamic process of self-referential inference whereby a system maintains representations of both its current state and overall goals. Bayesian networks are identified as a mathematically well-founded method for learning, representing, and inferring ontological knowledge. In particular, the Bayesian process of "explaining away" is an effective and principled way of integrating and jointly disambiguating evidence from a set of modalities to determine the most likely state of a given entity without a need for ad-hoc thresholds. The notion of reliability of observations, the integration of prior beliefs and observations, and a focus of attention mechanism allow this to be done efficiently and in a scalable fashion.

Acknowledgements

The author would like to acknowledge financial support from AT&T Labs Research and the Royal Commission for the Exhibition of 1851.

References

1. Abella, A.: From imagery to salience: Locative expressions in context. Ph.D. thesis, University of Columbia (1995)

2. Abella, A., Kender, J.: From pictures to words: Generating locative descriptions of objects in an image. ARPA94 pp. II:909–918 (1994)
3. Barnard, K., Duygulu, P., Forsyth, D.: Clustering art. In: Proc. Conference on Computer Vision and Pattern Recognition (2001)
4. Barnard, K., Duygulu, P., Forsyth, D., Freitas, N., Blei, D., Jordan, M.: Matching words and pictures. Journal of Machine Learning Research **3**, 1107–1135 (2003)
5. Barnard, K., Forsyth, D.: Learning the semantics of words and pictures. In: Proc. International Conference on Computer Vision (2001)
6. Bobick, A., Richards, W.: Classifying objects from visual information. Tech. rep., MIT AI Lab (1986)
7. Bunke, H., Pasche, D.: Structural Pattern Analysis, chap. Parsing multivalued strings and its application to image and waveform recognition. World Scientific Publishing (1990)
8. Buxton, H., Walker, N.: Query based visual analysis: Spatio-temporal reasoning in computer vision. Vision Computing **6**(4), 247–254 (1988)
9. Chen, Y., Rui, Y., Huang, T.: JPDAF based HMM for real-time contour tracking. In: Proc. Conference on Computer Vision and Pattern Recognition (2001)
10. Chua, T.S., Teo, K.C., Ooi, B.C., Tan, K.L.: Using domain knowledge in querying image databases. In: Proc. Int. Conference on Multimedia Modeling (1996)
11. Cooper, G., Herskovits, E.: A Bayesian method for the induction of probabilistic networks from data. Machine Learning **9**, 309–347 (1992)
12. Crowley, J., Coutaz, J., Rey, G., Reignier, P.: Perceptual components for context aware computing. In: Proc. Ubicomp 2002 (2002)
13. Darrell, T., Gordon, G., Harville, M., Woodfill, J.: Integrated person tracking using stereo, color, and pattern detection. In: Proc. Conference on Computer Vision and Pattern Recognition (1998)
14. Dennett, D.: Minds, machines, and evolution, pp. 129–151. Cambridge University Press (1984)
15. Duygulu, P., Barnard, K., De Freitas, J., Forsyth, D.: Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In: Proc. European Conference on Computer Vision (2002)
16. Ekin, A., Tekalp, A., Mehrotra, R.: Semantic video querying using an integrated semantic-syntactic model. In: Proc. International Conference on Image Processing (2002)
17. Friedman, N., Koller, D.: Being Bayesian about network structure. In: Proc. Conference on Uncertainty in Artificial Intelligence (2000)
18. Glöckner, I., Knoll, A.: Fuzzy quantifiers for processing natural-language queries in content-based multimedia. Tech. Rep. TR97-05, Faculty of Technology, University of Bielefeld, Germany (1997)
19. Guarino, N., Masolo, C., Vetere, G.: Ontoseek: Content-based access to the web. IEEE Intelligent Systems **14(3)**, 70–80 (1999)
20. Harnad, S.: The symbol grounding problem. Physica D **42**, 335–346 (1990)
21. Heckerman, D.: A tutorial on learning with Bayesian networks. In: M. Jordan (ed.) Learning in Graphical Models. MIT Press (1998)
22. Hongeng, S., Nevatia, R.: Large-scale event detection using semi-hidden markov models. In: Proc. International Conference on Computer Vision (2003)
23. Hoogs, A., Rittscher, J., Stein, G., Schmiederer, J.: Video content annotation using visual analysis and large semantic knowledgebase. In: Proc. Conference on Computer Vision and Pattern Recognition (2003)
24. Hu, M.: Visual pattern recognition by moment invariants. IRA Transactions on Information Theory **17**(2), 179–187 (1962)
25. Jaimes, A., Chang, S.: A conceptual framework for indexing visual information at multiple levels. In: IS&T SPIE Internet Imaging (2000)
26. Jaimes, A., Chang, S.F.: Integrating multiple classifiers in visual object detectors learned from user input. In: Proc. Asian Conference on Computer Vision (2000)
27. Jensen, F.: An Introduction to Bayesian Networks. Springer Verlag (1996)
28. Jordan, M. (ed.): Learning in Graphical Models. MIT Press (1999)
29. Katz, B., Lin, J., Stauffer, C., Grimson, E.: Answering questions about moving objects in surveillance videos. In: Proc of the AAAI Spring Symposium on New Directions in Question Answering (2003)
30. Kohler, C.: Selecting ghosts and queues from a car trackers output using a spatio-temporal query language. In: Proc. Conference on Computer Vision and Pattern Recognition (2004)
31. Kokar, M., Wang, J.: An example of using ontologies and symbolic information in automatic target recognition. In: Proc. SPIE Sensor Fusion: Architectures, Algorithms, and Applications VI, pp. 40–50 (2002)
32. Kruschwitz, U.: Exploiting structure for intelligent web search. In: Proc. Int. Conference on System Sciences. Maui, Hawaii (2001)
33. Lalmas, M.: Applications of Uncertainty Formalisms, chap. Information retrieval and Dempster-Shafer's theory of evidence, pp. 157–177. Springer (1998)
34. Lim, J.: Learnable visual keywords for image classification. In: Proc. ACM Int. Conference on Digital Libraries (1999)
35. Mezaris, V., Kompatsiaris, I., Strintzis, M.: An ontology approach to object-based image retrieval. In: Proc. International Conference on Image Processing (2003)
36. Miller, G., Beckwith, R., Fellbaum, C., Gross, D., Miller, K.: Introduction to Wordnet: an on-line lexical database. International Journal of Lexicography **3**, 235–244 (1990)
37. Mojsilovic, A., Gomes, J., Rogowitz, B.: Isee: Perceptual features for image library navigation. In: Proc. 2002 SPIE Human Vision and Electronic Imaging (2002)
38. Mueller, H., Marchand-Maillet, S., Pun, T.: The truth about Corel - evaluation in image retrieval. In: Proc. Conference on Image and Video Retrieval, LNCS 2383, pp. 38–50. Springer (2002)
39. Mueller, H., Mueller, W., Squire, D., Marchand-Maillet, S., Pun, T.: Performance evaluation in content-based image retrieval: Overview and proposals. Pattern Recognition Letters **22**(5), 593–601 (2001)
40. Murphy, K.: The Bayes Net Toolbox for Matlab. Computing Science and Statistics **33** (2001)
41. Nepal, S., Ramakrishna, M., Thom, J.: A fuzzy object query language (FOQL) for image databases. In: Proc. Int. Conference on Database Systems for Advanced Applications (1999)
42. Nevatia, R., Hobbs, J., Bolles, B.: An ontology for video event representation. In: Proc. Int. Workshop on Detection and Recognition of Events in Video (at CVPR04) (2004)
43. Nevatia, R., Zhao, T., Hongeng, S.: Hierarchical language-based representation of events in video streams. In: Proc. IEEE Workshop on Event Mining (2003)
44. Park, S., Aggarwal, J.: Event semantics in two-person interactions. In: Proc. Int. Conference on Pattern Recognition (2004)
45. Parsons, S., Hunter, A.: Applications of Uncertainty Formalisms, chap. A review of uncertainty handling formalisms, pp. 8–37. Springer (1998)
46. Pastra, K., Saggion, H., Wilks, Y.: Extracting relational facts for indexing and retrieval of crime-scene photographs. IEEE Intelligent Systems **18(1)**, 55–61 (2002)

47. Pfeffer, A., Koller, D.: Semantics and inference for recursive probability models. In: Proc. AAAI'00 (2000)
48. Pfeffer, A., Koller, D., Milch, B., Takusagawa, K.: SPOOK: A system for probabilistic object-oriented knowledge representation. In: Proc. Conference on Uncertainty in AI (1999)
49. Rodden, K.: Evaluating similarity-based visualisations as interfaces for image browsing. Ph.D. thesis, Cambridge University Computer Laboratory (2001)
50. Rowe, N., Frew, B.: Automatic classification of objects in captioned descriptive photographs for retrieval, chap. 4, pp. 65–79. AAAI Press (1997)
51. Roweis, S., Ghahramani, Z.: A unifying review of linear Gaussian models. Neural Computation **11(2)**, 305–345 (1999)
52. Roy, D.: Learning visually grounded words and syntax of natural spoken language. Evolution of Communication **4** (2001)
53. Roy, D.: A trainable visually-grounded spoken language generation system. In: Proc. Int. Conference of Spoken Language Processing (2002)
54. Sherrah, J., Gong, S.: Tracking discontinuous motion using Bayesian inference. In: Proc. European Conference on Computer Vision, pp. 150–166 (2000)
55. Sherrah, J., Gong, S.: Continuous global evidence-based Bayesian modality fusion for simultaneous tracking of multiple objects. In: Proc. International Conference on Computer Vision (2001)
56. Sinclair, D.: Voronoi seeded colour image segmentation. Tech. Rep. TR99-04, AT&T Laboratories Cambridge (1999)
57. Sinclair, D.: Smooth region structure: folds, domes, bowls, ridges, valleys and slopes. In: Proc. Conference on Computer Vision and Pattern Recognition, pp. 389–394 (2000)
58. Smith, P.: Edge-based motion segmentation. Ph.D. thesis, Cambridge University Engineering Department (2001)
59. Socher, G., Sagerer, G., Perona, P.: Bayesian reasoning on qualitative descriptions from images and speech. Image and Vision Computing **18(2)**, 155–172 (2000)
60. Spengler, M., Schiele, B.: Towards robust multi-cue integration for visual tracking. Lecture Notes in Computer Science **2095**, 93–106 (2001)
61. Town, C.: Ontology based visual information processing. Ph.D. thesis, University of Cambridge (2004)
62. Town, C.: Ontology-driven Bayesian networks for dynamic scene understanding. In: Proc. Int. Workshop on Detection and Recognition of Events in Video (at CVPR04) (2004)
63. Town, C., Sinclair, D.: Content based image retrieval using semantic visual categories. Tech. Rep. MV01-211, Society for Manufacturing Engineers (2001)
64. Town, C., Sinclair, D.: Ontological query language for content based image retrieval. In: Proc. IEEE Workshop on Content-based Access of Image and Video Libraries, pp. 75–81 (2001)
65. Town, C., Sinclair, D.: Language-based querying of image collections on the basis of an extensible ontology. International Journal of Image and Vision Computing **22(3)**, 251–267 (2004)
66. Tsai, W., Fu, K.: Attributed grammars - a tool for combining syntactic and statistical approaches to pattern recognition. IEEE Transactions on Systems, Man and Cybernetics **SMC-10**(12) (1980)
67. Tsotsos, J., Mylopoulos, J., Covvey, H., Zucker, S.: A framework for visual motion understanding. IEEE Trans. on Pattern Analysis and Machine Intelligence **Special Issue on Computer Analysis of Time-Varying Imagery**, 563–573 (1980)
68. Wachsmuth, S., Socher, G., Brandt-Pook, H., Kummert, F., Sagerer, G.: Integration of vision and speech understanding using Bayesian networks. Videre: A Journal of Computer Vision Research **1**(4) (2000)
69. Wu, Y., Huang, T.: A co-inference approach to robust visual tracking. In: Proc. International Conference on Computer Vision (2001)
70. Xu, C., Prince, J.: Snakes, shapes, and gradient vector flow. IEEE Transactions on Image Processing **7**(3), 359–369 (1998)
71. Zhao, R., Grosky, W.: From features to semantics: Some preliminary results. In: Proc. IEEE Int. Conference on Multimedia and Expo, pp. 679–682 (2000)