Ontology based Visual Information Processing

Christopher Phillip Town

Trinity College

University of Cambridge
Computer Laboratory

A thesis submitted for the degree of *Doctor of Philosophy* at the University of Cambridge

December 2004

Abstract

Despite the dramatic growth of digital image and video data in recent years, many challenges remain in enabling computers to interpret visual content. This thesis addresses the problem of how high-level representations of visual data may be automatically derived by integrating different kinds of evidence and incorporating prior knowledge. A central contribution of the thesis is the development and realisation of an inference framework for computer vision founded on the application of ontologies and ontological languages as a novel methodology for active knowledge representation. Three applications of the approach are presented:

The first describes a novel query paradigm called OQUEL (ontological query language) for the content-based retrieval of images. The language is based on an extensible ontology which encompasses both high-level and low-level visual properties and relations. Query sentences are prescriptions of desired image content in terms of English language words.

The second application extends the notion of ontological query languages to the field of automatic event detection from video footage. It is shown how the structure and parameters of Bayesian networks can be learned to automatically generate effective high-level state and event recognition mechanisms from an ontology consisting of a ground truth schema and a set of annotated training sequences. Recognition is improved by augmenting the original ontology with automatically extracted visual content descriptors.

The third example considers a sensor fusion problem in which computer vision information obtained from calibrated cameras is integrated with a sentient computing system. Fusion of the different sources of information relies on Bayesian networks to model dependencies and reliabilities of the multi-modal variables. The system maintains a world model which serves as a domain ontology and incorporates aspects of both the static and dynamic environment.

Declaration of Originality

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration, except where specifically indicated in the text. It does not contain material that has already been or is currently being submitted for any degree or qualification, nor does it exceed the word limit specified by the Computer Laboratory Degree Committee.

The following are the principal publications derived from this work:

- Town, C.P. *Vision-based Augmentation of a Sentient Computing World Model*, International Conference on Pattern Recognition, 2004

- Town, C.P. *Ontology-driven Bayesian Networks for Dynamic Scene Understanding*, International Workshop on Detection and Recognition of Events in Video (at CVPR04), 2004

- Town, C.P. and Moran, S.J. *Robust Fusion of Colour Appearance Models for Object Tracking*, British Machine Vision Conference, 2004

- Town, C.P. and Pugh, D.J. *Combining Contour, Edge and Blob Tracking*, International Conference on Image Analysis and Recognition, 2004

- Town, C.P. *Ontology-guided Training of Bayesian Networks for High-level Analysis in Visual Surveillance*, Proc. IEEE International Workshop on Performance Evaluation in Tracking and Surveillance (at ECCV04), 2004

- Town, C.P. and Sinclair, D.A. *Language-based Querying of Image Collections on the basis of an Extensible Ontology*, International Journal of Image and Vision Computing, vol. 22(3), 2004

- Town, C.P., *Adaptive Integration of Visual Tracking Modalities for Sentient Computing*, International Workshop on Visual Surveillance and Performance Evaluation in Tracking and Surveillance, 2003

- Town, C.P., *Computer architecture for self-referential perceiving systems*, Perception Journal special supplement (Proceedings of the European Conference on Visual Perception), vol. 32(1), 2003

- Town, C.P. and Sinclair, D.A. *A Self-referential Perceptual Inference Framework for Video Interpretation*, International Conference on Vision Systems, 2003, and Lecture Notes in Computer Science, Springer, Volume 2626, pp. 54-67

- Town, C.P. *Goal-directed Visual Inference for Multi-modal Analysis and Fusion*, IEE Conference on Visual Information Engineering, 2003

Christopher Phillip Town, Trinity College

Date

Acknowledgements

I would like to thank Dr John Daugman and Dr David Sinclair for their guidance and encouragement. Thanks are also due to Dr David Belanger and Prof. Andy Hopper for their support, and to the Cambridge University Computer Laboratory for providing facilities.

I also gratefully acknowledge financial assistance from the Royal Commission for the Exhibition of 1851 and AT&T Labs Research. Trinity College Cambridge provided me with additional support and a stimulating collegiate environment.

Dedication

This thesis is dedicated to my girlfriend, Suvina Jayatilaka, and to my parents, Dr Michael-Harold Town and Sabine Town. Without their love and support, this work would not have been possible.

In loving memory of my grandfather, Harold Town (22nd Dec 1914 - 26th Feb 2005).

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview and Contributions

This thesis discusses the design and implementation of computational models for artificial perceiving systems. It addresses the challenging problem of how high-level descriptions of visual information may be automatically derived from different kinds of evidence. It argues that these goals are achievable by exploiting mutual interaction between different levels of representation and integrating different sources of information both horizontally (multi-modal and temporal fusion) and vertically (bottom-up, top-down) by incorporating prior knowledge. In particular, this thesis shows how these aims can be achieved through the appropriate use 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.

A central contribution of this thesis is the development and realisation of an inference framework for computer vision founded on the application of ontologies as a novel active knowledge representation methodology. This representation allows one to pose visual analysis problems in terms of queries expressed in a visual language incorporating prior hierarchical knowledge of the syntactic and semantic structure of entities, relationships, and events of interest. Perceptual inference then takes place within an ontological domain defined by the structure of the problem and the current goal set. Moreover, this thesis argues that an extensible framework for general and robust perceptual inference is best attained by pursuing an inherently flexible "self-referential" approach. Such a system embodies an explicit representation of its own internal state and goals.

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. In this thesis, a family of methods which has gained prominence in recent years, namely Bayesian networks, is used extensively. Formally, these are graphical models which represent a factorised joint distribution over a set of random variables by encoding the (in)dependence relations between them. The structure and implicit constraints of such networks allows one to represent knowledge of entities, properties and their relationships. Statistical and machine learning techniques are applied to acquire the parameters of such models and various inference methods are carried out to perform high-level analysis of the model and its temporal evolution. 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.

Whereas speech and language processing techniques are concerned with the analysis of sound patterns, phonemes, words, sentences, and dialogues, video and image analysis is confronted with pixels, video frames, primitive features, regions, objects, motions, and events. An important difference [254] between the two arises from the fact that visual information is inherently more ambiguous and semantically impoverished. The rules underlying the definition of concepts such as scenes and events in terms of the observed primitives are thus even more difficult to model and implement on a computer system than is the case with natural language. There consequently exists a wide semantic gap between human interpretations of image information and those currently derivable by a computer.

The research presented in this thesis demonstrates how this gap can be narrowed by means of an ontological language that encompasses a hierarchical representation of task-specific attributes, objects, relations, temporal events, etc., and relates these to the processing modules available for their detection and recognition from the underlying medium. Words in the language therefore carry meaning directly related to the appearance of real world objects. Tasks such as retrieval of images matching a set of user criteria, automated visual surveillance of a scene, and visually mediated human computer interaction can then be carried out by processing sentence structures in an appropriate ontological language. 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 that permits incremental refinement of alternate hypotheses through the fusion of multiple sources of information and goal-directed feedback to facilitate disambiguation in a context specified by the current set of ontological sentences.

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 that are of value given a particular task are evaluated and evaluation may stop as soon as the appropriate top level symbol sequence has been generated.

This thesis presents concrete implementations of the approach discussed above which demonstrate its utility for solving relevant research problems. Three main applications are described which draw on work in computer vision, machine learning, probabilistic reasoning, knowledge representation, information retrieval, human computer interaction, and ubiquitous computing. Avenues for further work in these and other domains are discussed at the end of the respective chapters and in chapter 7.

1.2 Research Outline

1.2.1 Content-based Image Retrieval

The first application of the research consists of an image retrieval system which allows users to search image databases by posing queries over desired visual content of images that the system is to retrieve from a large image collection. A novel query paradigm 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 image properties and relations. Query sentences are prescriptions of target image content rather than exhaustive descriptions. They can represent abstract and arbitrarily complex retrieval requirements at different levels of conceptual granularity.

The retrieval process takes place entirely within the ontological domain defined by the syntax and semantics of the user query and integrates multiple sources of evidence. It utilises automatically extracted image segmentation and classification information, as well as Bayesian networks to infer higher level and composite terms. Since the system does not rely on the pre-annotation of images with sentences in the language, the format of text queries is highly flexible. The language is also extensible to allow for the definition of higher level terms such as "cars", "people", etc. on the basis of existing language constructs through the use of Bayesian inference networks.

This thesis argues that the OQUEL language provides an effective mechanism for addressing key problems of content-based image retrieval, namely the ambiguity of image content and user intention and the semantic gap that exists between user and system notions of relevance. By basing such a language on an extensible ontology, one can explicitly state ontological commitments about categories, objects, attributes, and relations without having to pre-define any particular method of query evaluation or image interpretation. 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.

1.2.2 Dynamic Scene Understanding

The second application example presents an extension of the notion of ontological query languages to the field of automatic event detection from video footage. It shows how modern techniques for structure and parameter learning in Bayesian networks can be used to automatically generate effective high-level state and event recognition mechanisms from an ontology consisting of a ground truth schema and a set of annotated training sequences. It is demonstrated how the performance of such inference networks can be improved by augmenting and grounding the original ontology with additional types of visual content extraction and representation such as state-of-the-art object detection, appearance modelling, and tracking methods.

The integration of these different sources of evidence is optimised with reference to the syntactic and semantic constraints of the ontology. By applying these techniques to a visual surveillance problem, it is shown how high-level event, object and scenario properties may be inferred on the basis of the visual content extraction modules and an ontology of states, roles, situations and scenarios that is derived from a pre-defined ground truth schema. Provided sufficient amounts of data are available, this process exhibits robustness to both errors in the ground truth and failures of the automatically extracted visual content descriptions. Furthermore, performance analysis of the resulting tracking and analysis framework provides a useful basis for comparison of alternative schemas and methods. It also allows alternative ontologies to be compared for their self-consistency and realisability in terms of different visual detection and tracking modules.

1.2.3 Sentient Computing

The third example considers a sensor fusion problem in which computer vision information obtained from calibrated cameras is integrated with a sentient computing system known as "SPIRIT", which employs an ultrasonic location infrastructure to track people and devices in an office building. Sentient computing aims to model aspects of the context within which human-computer interactions take place in order to better infer and anticipate user intentions and requirements. This is achieved by integrating information from a range of sensors distributed throughout an office environment in order to maintain an internal representation, or world model, of that environment.

The thesis reports on research which fuses a range of computer vision modules with the SPIRIT system. Vision techniques include background and object appearance modelling, face detection, segmentation, and tracking modules. Integration is achieved at the system level through the metaphor of shared perceptions in the sense that the different modalities are guided by and provide updates for a shared internal model. This world model incorporates aspects of both the static (e.g. positions of office walls and doors) and dynamic (e.g. location and appearance of devices and people) environments. It serves both as an ontology of prior information and as a source of context that is shared between applications. Fusion and inference are performed by Bayesian networks that model the probabilistic dependencies and reliabilities of different sources of information over time. It is shown that the fusion process significantly enhances the capabilities and robustness of both modalities, thus enabling the system to maintain a richer and more accurate world model.

1.3 Dissertation Structure

- *Chapter 2*: This chapter presents a survey of published research related to this thesis.

- *Chapter 3*: This introduces some of the key concepts and the overall approach to hierarchical knowledge representation by means of ontologies. It is shown how visual processing may then be regarded as parsing sentences expressed in a language defined by reference to an ontology that characterises a particular domain and task of interest.

- *Chapter 4*: This chapter illustrates and applies the key concepts of this thesis through the design of an ontological query language for content-based image retrieval. It is shown how such a language can serve as a means of bridging the semantic gap between image indexing systems and their users.

- *Chapter 5*: An extension of the ontological framework to dynamic scene analysis from video is described. The mechanisms required to ground ontological descriptors given a set of visual processing modalities and a domain description for automated video surveillance are presented.

- *Chapter 6*: This chapter describes how the methods introduced thus far can be applied to achieve multi-sensory and multi-modal fusion for a sentient computing system. Information from visual and non-visual modalities is integrated through a central world model that serves as a shared ontology representing aspects of the environment.

- *Chapter 7*: The final chapter summarises the main achievements of the thesis, discusses results obtained thus far, and provides suggestions for further work.

Chapter 2

Survey of Related Research

This chapter examines some relevant research trends and presents an overview of published work that is related to this thesis. A number of additional references to related work are given in the following chapters.

2.1 Relevant Trends in Visual Recognition and Machine Learning

A lot of the research referenced below can be broadly viewed as following the emerging trend that object recognition and the recognition of temporal events are best approached in terms of general language processing which attempts a machine translation [70] from information in the visual domain to symbols and strings composed of predicates, objects, and relations. The general idea is that recognising an object or event requires one to relate ill-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. Many state-of-the-art recognition systems therefore explicitly or implicitly employ a probabilistic grammar defining syntactic rules for recognising compound objects or events based on the detection of individual components corresponding to detected features in time and space. Recognition 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 idea has a relatively long heritage in syntactic approaches to pattern recognition [276, 32]. Interest has been revived recently in the image and video analysis community following the popularity and success of probabilistic methods such as Hidden Markov models (HMM) and related approaches adopted from speech and language processing research [215].

While this approach has shown great promise for applications ranging from image retrieval to face detection and visual surveillance, a number of problems remain. 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 [153, 82] continues to grow. Recently there has been a very strong trend towards using Bayesian techniques for learning and inference, especially factorised graphical probabilistic models [135, 141] such as Bayesian networks. Such methods provide a natural tool for dealing with the most common problems of engineering and science, namely complexity and uncertainty. They have recently become predominant in a wide variety of fields ranging from control engineering, speech processing, and data mining, to decision support, biotechnology, and medical image classification.

Probabilistic graphical models [135, 107] incorporate prior information about conditional independences amongst a set of variables corresponding to observations and hidden causes that are to be inferred using inference techniques. Dynamic Bayesian networks [85] embody a stochastic state that is dynamically adapted to arrive at the most likely model ("belief") of the world, i.e. the conclusion that is best supported by the available data. Formally, they are annotated directed acyclic graphs that encode a factored joint probability distribution over a set of variables and can model the evolution of probabilistic state over time. They generalise [226] other popular multivariate approaches such as HMMs, factor analysis, Kalman filters, and Ising models. The parameters of such a model can be learned from training data using techniques such as Expectation Maximisation (EM) or gradient descent.

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, because the underlying causal links and prior conditionalisation are made more explicit. The recent development of various approximation schemes [164, 91] 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) cues to disambiguate and combine evidence in the most suitable context as defined by the top level model structure.

A parallel trend [82] in machine learning concerns the advance of statistical learning theory, particularly the use of Support Vector machines (SVM) [280]. These methods are founded on the concept of structural risk minimisation as a method for learning a parameterised approximation of a target function by minimising an empirical loss criterion subject to smoothness or complexity constraints. The learned function is represented in terms of kernel basis functions, which are generally viewed as computing dot products between input vectors in an induced feature space. An appropriate choice of the kernel mapping allows even highly complicated decision problems to be represented by means of linear classification boundaries (hyperplanes) in the feature space. Minimising the structural risk can then be reduced to solving geometric constraints to find boundaries that maximise the margin between kernel-mapped input vectors that are assigned different class labels. The representation of the boundaries through a small set of so-called support vectors allows very efficient classifiers to be learned from data. Numerous extensions to the original SVM formalism have been proposed in recent years [266] and are finding increasing application in computer vision tasks [111].

Computer vision researchers have also been turning to a number of other machine learning techniques such as boosting methods for multi-classifier combination [265, 283], bootstrapping [273], and reinforcement learning [201].

An open debate in object recognition and tracking research concerns the extent to which methods should be made to rely on explicit models (object-based, parametric) as opposed to exemplars (view-based, largely non-parametric) which are selected instances of the target object observed in the training data. Models are harder to construct but allow more concise representations than exemplars, which become very complex as one tries to adequately capture the variations in the modelled structure. There is a trend towards schemes that acquire (learn) models from data and hybrid schemes that combine explicit model representations with view-based frameworks [146].

Devising a near optimal strategy for recognising objects by applying the most appropriate combination of visual routines such as segmentation and classification modules can be learned from data. For example, recent work on adaptive object recognition [68] has shown how dynamic strategies for high-level vision can be acquired by modelling the recognition problem as a Markov decision process.

For a general overview of recent work in relevant areas, [8], [88], and [177] offer surveys of human motion analysis, [42] gives more recent references to object detection and tracking research, and [295] presents a review of work in vision-based gesture recognition. A survey of some earlier work on forming correspondences between linguistic and visual information was given in [257].

2.2 Linking Language to Visual Data

Among systems for generating natural language like descriptions from video, the ambitious VITRA (VIsualTRAnslator, [286, 114, 113]) project aimed at an incremental analysis of video sequences to produce simultaneous narration. VITRA was used to incrementally generate reports about real-world traffic scenes or short video clips of football matches using an integrated knowledge-based system capable of translating visual information into natural language descriptions.

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. By representing spatial relationships using inference networks, logic minimisation techniques can be applied to select the shortest unambiguous description from a spanning tree of possible descriptions.

Learning associations between visual keywords and image properties is of particular interest for content-based image retrieval [225, 159, 299, 269], 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 [134], whereby a user defines visual object models via an object-definition hierarchy (region, perceptual-area, object part, and object).

Recent work [16, 14] 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, [226]) 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 [175] 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.

2.3 Video Analysis and Tracking

Bayesian techniques [135, 107] have been heavily used in video content analysis and target tracking in recent years [33]. They offer a way of constructing richer target and noise models which improve tracking and feature extraction performance in cluttered and dynamic environments. At the same time such models can be factorised to reduce their complexity by making use of the causal dependence and independence relations that can be applied to different sources of visual evidence. While the computational requirements of such approaches are higher than those of simpler blob trackers [294], the increasing sophistication of the algorithms coupled with continued exponential growth of computational power are driving their advance.

Much recent work in visual tracking has focused on particle filtering methods such as Condensation (conditional density propagation) [128, 129] for tracking. These are sequential Markov Chain Monte Carlo (MCMC) techniques based on weighted sequential stochastic sampling by means of a set of "particles" which approximates a potentially very complicated (multi-modal) probability distribution. In practice they can suffer from robustness and data association problems, particularly as the dimensionality of the state space increases. The number of required particles is difficult to predict, especially for high-dimensional distributions, although more efficient schemes have been proposed recently which address some of these issues [169, 193].

One of the earlier examples of using Dynamic Bayesian networks for visual surveillance appears in [34]. DBNs offer many advantages for tracking tasks such as incorporation of prior knowledge and good modelling ability to represent the dynamic dependencies between parameters involved in a visual interpretation.

In [241] tracking of a person's head and hands is performed using a Bayesian network which deduces the body part positions by fusing colour, motion and coarse intensity measurements with context dependent semantics. It is shown that the system does not rely on continuity assumptions and offers some robustness to occlusions.

Later work by the same authors [242] 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 ("explaining away"). This makes it possible to track multiple objects with self-starting hypotheses, although some performance issues remain to be resolved. The principle of Bayesian Modality Fusion (BMF) had been introduced earlier by [272], although their work was limited to discrete spatial variables and uni-modal observations. An adaptive variant of BMF is presented in [256] which shows how the robustness of face tracking using multiple cues can be enhanced by using the Condensation [128] method to maintain multiple target hypotheses.

An approach to visual tracking based on co-inference of multiple modalities is also presented in [296]. A Sequential Monte Carlo approach is applied to co-infer target object colour, shape, and position. The method uses variational analysis to yield a factorised probabilistic model which is of lower dimensionality than those utilised by comparable approaches such as Condensation. The implementation works in real-time but is limited to tracking a single non-articulated object at present.

An application of DBNs for multi-modal fusion of visual cues in human computer interaction is given in [217], which considers the task of detecting when a user is speaking. The model combines inputs from vision modules that detect faces, skin colour, skin texture, and mouth motions.

Tracking objects based on combination of several cues can also be achieved using HMMs. In [46] 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. The authors report real time performance for a contour-based object tracking task.

When accurate tracking measurements are available, e.g. when tracking light points attached to known locations on the human body, even fairly subtle tasks such as gender recognition on the basis of gait and other movement cues can be performed automatically using techniques such as neural networks and decision trees [233]. Such work has connections to earlier results in psychophysics [57, 58] that showed how people could identify complex human motions purely on the basis of moving light displays.

Much recent tracking work addresses the challenge of combining data-driven exemplar-based and model-based methods. Interesting work by Toyama et al. [271] presents an approach that focuses on using exemplars to train a probabilistic generative contour model of tracked objects and a non-generative noise model. “Metric mixture” models are learned from data to represent the exemplars. Tracking takes place in a metric space with motion dynamics represented by a Markov model.

In a similar vein, [136] features the EM-based learning of robust adaptive appearance models for motion tracking of natural objects. The approach involves a mixture of stable image structure, learned over long time courses, along with 2-frame motion information and an outlier process. An online EM algorithm is used to adapt the appearance model parameters over time, which makes the tracking robust with respect to reasonably gradual view changes and occlusions.

[140] describes an effective approach to learning exemplar-based representations called “flexible sprites” in a layered representation of video. A variational EM algorithm is used to learn a mixture of sprites from a video sequence by performing probabilistic inference on each frame to infer the sprite class, translation, and mask values, including occlusions between layers. Currently the method suffers from some drawbacks in that the number of sprites needs to be fixed in advance and the layering must not change during the sequence.

[208] presents a modular approach which combines several simple target detection methods using a Kalman tracker. Targets are represented by regions of interest whose spatial extent is defined by 2D Gaussian functions. The background image is slowly updated over time to factor in lighting changes. Target groups are formed and split using a Mahalanobis distance threshold but target identity is not preserved. Currently there are only two simple detection modules (based on motion history and background differencing) but the underlying architecture is promising and yielded good results on test sequences.

[238] shows how even very simple knowledge of people shape and interactions can be integrated in a tracking application to make it more robust to partial or temporarily complete occlusions.

[75] uses a relatively sophisticated mixture of Gaussians model for the background and a Bayesian network to reason about the state of tracked objects. Objects are represented by a combination of predictive features such as position, size, colour, and occlusion status as predicted using a scene model and Bayesian network.

Tracking and object recognition using mixtures of tree-structured probabilistic models is discussed in [124, 125]. The models represent the appearance of local features and geometric relationships between object parts for different viewing conditions. Models can be automatically learned from training data and incorporate a large number of features to exploit redundancies in order to make matching robust to feature detection errors and occlusions. The underlying tree structure allows efficient correspondence search using dynamic programming techniques. Recently it was shown [216] how the mixture model can be simplified to a simple tree model to allow more efficient detection and tracking of objects.

Bayesian networks also play an important role in content-based indexing of video. For example, [281] considers indexing of a movie database by exploiting characteristic structure. In [72] extraction of semantic events from video is again achieved using Bayesian networks based on a hierarchy of object motions and object interactions. This allows searching in video archives based on a query-by-example paradigm for action descriptions.

In [45] simple object segmentation and tracking techniques as well as compressed MPEG block motion features are used to index video for retrieval purposes. Queries can be formulated using a diagrammatic representation of desired object motions with optional inclusion of keyword strings and feature (colour, shape, size) constraints.

Scene change detection for video indexing can also be performed without tracking through frame activity analysis in terms of temporal and spatial displacements of pixel blocks [95]. Similarly, video event recognition has been demonstrated through learning of temporal change signatures at the pixel level without the need for tracking or trajectory modelling [93]. High-level behavioural analysis of human motions can then be performed by integrating this information using a Bayesian network without the need for explicit object segmentation.

[59] presents a system which tracks users by combining information from a stereo camera system with skin colour modelling and face detection. Tracking amounts to finding correspondences for the most reliable modalities, while face pattern comparisons allow the system to handle occlusions and object re-appearances.

Tracking on the basis of multiple sources of information is also demonstrated in [47], which presents a system for fusing auditory and visual cues for speaker detection. As an improvement to earlier work by the same authors, the fusion is performed by a Dynamic Bayesian network whose structure was learned by means of a modified AdaBoost [236] algorithm. Their earlier work [199] also demonstrated fusion of visual and auditory modalities (face, skin, texture, mouth motion, and silence detectors) by means of a DBN whose performance was improved through error feedback and boosting, but whose network structure was fixed *a priori*.

In addition to multi-modal fusion, information from multiple cameras can also offer valuable cues for tracking. [64] shows an approach to self-organising cue integration from multiple calibrated cameras by means of the Democratic Integration

[274] method.

A different approach to adaptation of improved observation models for visual tracking is presented in [131]. Colour models for objects tracked via a particle filter are adapted using Expectation Maximisation in a manner which takes account of appearance changes due to observed motion as opposed to static effects such as changed lighting conditions.

2.4 Recognition of Actions and Structured Events

Over the last 15 years there has been growing interest within the computer vision and machine learning communities in the problem of analysing and recognising human behaviour from video. Such systems typically consist of a low or mid level computer vision system to detect and segment a human being or object of interest, and a higher level interpretation module that classifies motions into atomic behaviours such as hand gestures or vehicle manoeuvres. Higher-level visual analysis of compound events has in recent years also been performed using parsing techniques and a probabilistic grammar formalism.

Considerable work has been carried out on building traffic surveillance systems capable of representing moving vehicles by a series of verbs or short descriptive sentences [184, 150, 185]. Such relatively early work usually relied on somewhat ad-hoc methods for translating parameterised motion models assigned to objects detected by template matching. Natural language descriptions are generated by a process of frame instantiation to fit pre-defined action scenarios. More recent methods are capable of recognising more complicated behavioural patterns, although they remain limited to fairly circumscribed scenarios such as sport events [113, 123], presentations [142], small area surveillance [264, 218, 194], and game playing [179].

[148] presents an ontology of actions represented as states and state transitions hierarchically organised from most general to most specific (atomic). The paper stresses the point that language-based description of video requires one to establish correspondences between concepts and features extracted from video images. Appropriate syntactic components such as verbs and object labels can then be determined to generate a natural language sentence from a video sequence.

In [209] the authors present a control architecture in which a so-called approximate world model represents the complex relationships among objects of interest. Visual processing can then be performed on the basis of qualitative terms relating only to those objects that are relevant for a given context. This idea is applied to the problem of autonomous control of cameras in a TV studio.

The role of attentional control for video analysis was already pointed out in [35]. The system described there performs selective processing in response to user queries for two biological imaging applications. This gives the system a goal directed attentional control mechanism by selecting the most appropriate visual analysis routines in order to process the user query.

Selective visual processing on the basis of Bayesian networks and decision theory has also been demonstrated in control tasks for active vision systems [219]. Knowledge representation using Bayesian networks and sequential decision making on the basis of expected cost and utility allow selective vision systems to take advantage of prior knowledge of a domain's abstract and geometric structure and the expected performance and cost of visual operators.

[166] stresses that the derivation of a qualitative model of the underlying dynamics is required for a system to have some understanding of the observed interactions of objects. The paper considers analysis in terms of Newtonian mechanics of a simplified scene model using a linear programming technique. Interpretations are expressed as assertions about the kinematic and dynamic properties of the scene relative to a preference hierarchy to select the most plausible interpretation of the observed interactions.

[200] describes a novel DBN-based switching linear dynamic system (SLDS) model and presents its application to figure motion analysis. A key feature of the approach is an approximate Viterbi inference technique to overcome the intractability of exact inference in mixed-state DBNs.

An interesting two-level approach to parsing actions and events in video is described in [25, 130]. HMMs are used to detect candidate low-level temporal features, which are then parsed using a stochastic context free grammar (SCFG) parsing scheme that adds disambiguation and robustness to the stream of detected atomic symbols. The resulting system is capable of generating a discrete symbol stream from continuous low-level detections, enforcing temporal exclusion constraints during parsing, and correcting certain types of detection errors based on the current parsing state.

A similar approach is taken by [179] which uses the Earley-Stolcke parsing algorithm for stochastic context-free grammars to determine the most likely semantic derivation for recognition of complex multi-tasked activities from a given video scenario. Symbolic parsing strategies based on methods first developed by [9] are introduced that are able to correct certain types of wrong symbol insertion, substitution, and deletion errors to make the recognition much more robust to detection errors and plan deviations. The approach is shown to work well when detecting fairly complex behaviours in a multi-player card game.

Parsing techniques for multi-modal integration are also finding applications in human computer interface research. For example, [139] presents a system for speech recognition using a general probabilistic framework for multi-modal ambiguity resolution. This is achieved using a weighted finite-state device which takes speech and gesture streams as inputs and derives a joint interpretation. It is an extension to earlier work [138] by the authors which used a unification-based grammar in conjunction with a multi-dimensional chart parser.

A method for recognising complex multi-agent action is presented in [123]. Bayesian networks are again used to probabilistically represent and infer the goals of individual agents and integrate these over time from visual evidence. The method is applied to the automatic annotation of sport scenes in American Football by representing each formation of players by Bayesian networks based on visual evidence and temporal constraints. A similar agent-based approach was used in [218] for a surveillance application.

Bayesian techniques for integrating bottom-up information with top-down feedback have also been applied to challenging tasks involving the recognition of interactions between people in surveillance footage. In [194] Coupled Hidden Markov models (CHMM) are shown to give good results for learning interaction semantics from real and synthetic training data.

[260] applies a multi-hypothesis tracking system to acquire joint co-occurrence statistics of tracked objects and their properties (silhouette, position, and velocity) from a large quantity of surveillance footage. Based on this, hierarchical binary

tree classifiers are trained to detect patterns of activity which re-occur over several days.

The benefits of using sequence data for classifier training are also demonstrated by [300], which presents a method for face recognition from video sequences. The joint probability distribution of facial identity and appearance is represented by sequential importance sampling. Video-based face recognition is also the subject of [160], which uses HMMs trained from annotated sequences to match faces based on motion dynamics of features derived by PCA.

2.5 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 [277] 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 that 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 [49], 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).

[285, 252] 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, [227, 228] 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 [15] feasible, which matches still image annotations with region properties using hierarchical clustering and EM.

[198] describes an active vision system in which scene knowledge is represented through a combination of semantic networks and statistical object models. The methods are used to guide an active camera which performs object recognition and localisation tasks in an office room.

In [56], an architecture for perceptual computing is presented that 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 [149], 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 [188], 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 [187], 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.

[174] 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 [71], 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 [197], 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 [133]. They also play an important role in improving content-based indexing and access to textual documents (e.g. [98], [151]), where they can be used for semantics-based query expansion and document clustering.

[144] 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.

Chapter 3

Ontologies and Languages for Computer Vision

This chapter gives an account of some concepts and approaches in the general fields of ontology, language and vision. It highlights certain recurrent themes and emergent trends within these broad disciplines which motivate and underpin much of the central argument of this thesis. Section 3.2 describes the underlying ideas in more detail.

3.1 Current and Historical Formulations

Ontology is the theory of objects in terms of the criteria which allow one to distinguish between different types of objects and the relations, dependencies, and properties through which they may be described. Ontologies encode the relational structure of concepts which one can use to describe and reason about aspects of the world. They can play both a passive representational or taxonomic role in terms of what is said to exist in the world, and an active inferential role which defines how the relevant properties of the world may be identified.

This makes them eminently suitable to many problems in computer vision that require prior knowledge to be modelled and utilised in both a descriptive and prescriptive capacity. While some relevant work on the use of ontologies in computer vision was described in chapter 2, this section will provide a broader context.

3.1.1 Philosophical Origins

The study of ontology has a long lineage in philosophy where it concerns the fundamental metaphysical investigation of the nature of being in terms of the entities and categories which exist (or could exist). Its purpose is to capture and reason about the foundations of knowledge as opposed to its origins (which is the subject of *epistemology*). The term ontology was only popularised during the early 17th century [249], although the endeavour itself can be traced back to the classical philosophers such as Aristotle (384-322 BC). Both Aristotle and his former teacher Plato regarded philosophy as a search for universals. However, while Plato focused on universals (his "ideal forms") as having separate existence from their particular instances, Aristotle preferred to reason from these particulars to their abstract essence. One can see these thought processes as early examples of both deductive and inductive reasoning. Many similar and sometimes earlier accounts of ontology and epistemology exist in Eastern philosophies such as Chinese Taoism and the Indian *Vaisesika Sutra* ("Atomic Doctrine", ca. 500BC).

Philosophical ontology is thus primarily concerned with "an account of being in the abstract"[1], although in modern times it has increasingly taken a more empirical character that stresses the need for well-founded descriptions of reality.

3.1.2 Ontologies of Science

While the fundamental study of ontology seeks to establish universal and largely qualitative accounts of reality, science demands a more practical quantitative approach that aims to provide explanations and predictions of observable phenomena. Each branch of science has its own shared body of theories, terminology, and conventions, which represent the basic *ontological commitments* of the given discipline. For example, the principle concepts of chemistry include atoms (categorised into the elements), molecules, and reactions. Physics on the other hand explains natural phenomena using notions such as particles, waves, forces, energy, and the dimensions of space and time.

This view of science has motivated many influential philosophers of science such as Quine [214], who see science as the most reliable process of establishing

[1] This is also the first known English definition of *ontology* as it appeared in Nathan Bailey's *Dictionarium Britannicum* of 1721.

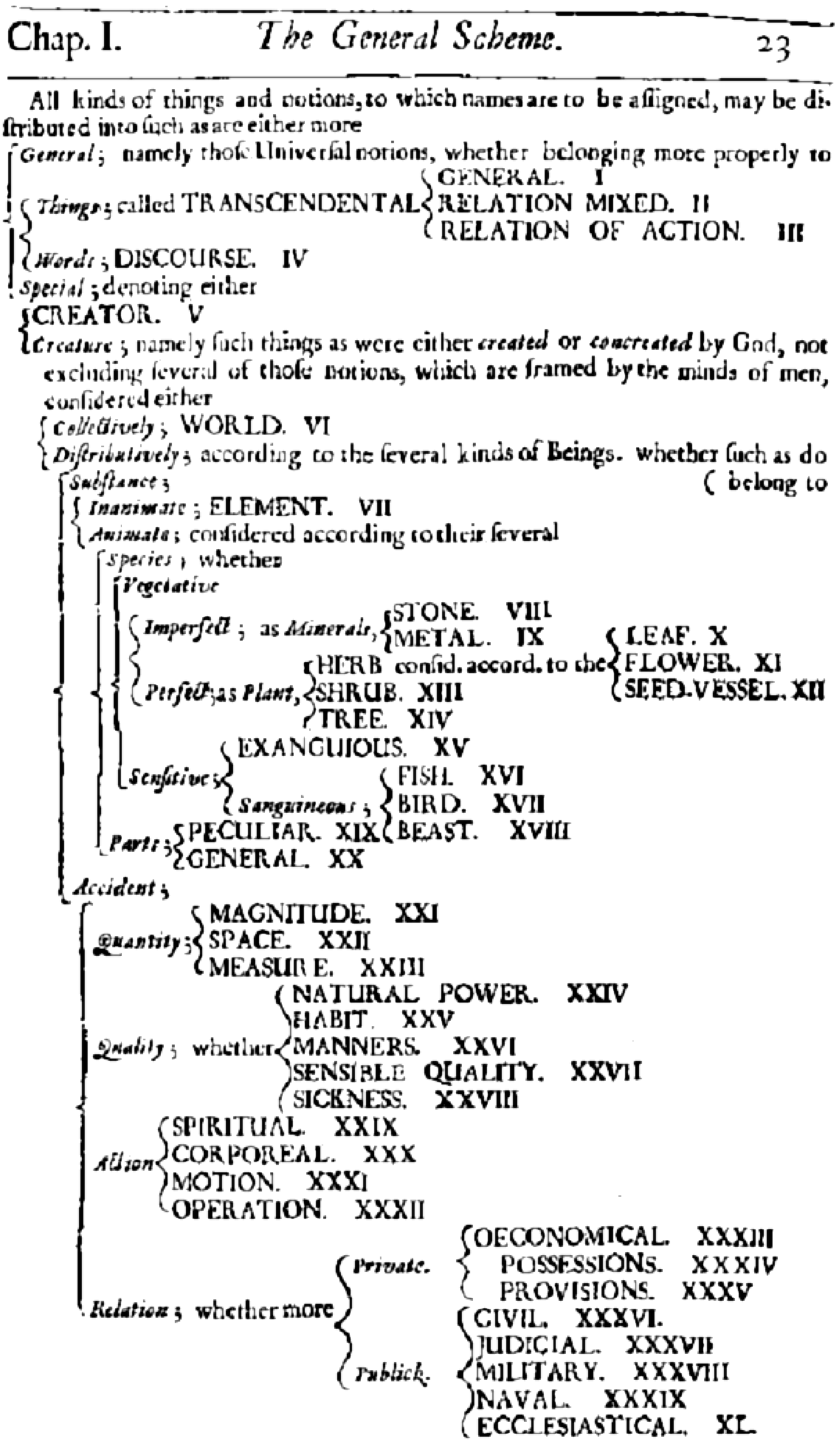
empirically verifiable (or rather falsifiable, as stressed by Popper [210]) theories about the physical universe at different levels of abstraction. Empirical ontology is therefore distinct from purely categorical ontology in that the former seeks to answer (and provide evidence for) the question “what is there” rather than “what could there be”.

Ontologies are often formalised in the natural sciences in a more restricted sense as taxonomies, i.e. as (typically hierarchical) classification schemes together with a set of matching criteria and an underlying theory. An early example is the seminal *Systema Naturae* written by Carolus Linnaeus (1707–1778) in which he laid out taxonomical and terminological conventions for biological organisms. These include the standard binomial nomenclature of genus and species name[2] and the hierarchical organisation of animals and plants according to kingdom, phylum/division, class, order, family, genus, and species. This scheme is still in use today, though the particular assignments are being revised by evolutionary and microbiological accounts of speciation. Other examples of ontologies in the natural sciences include the periodic table of the elements and the standard particle models of quantum physics. Applied sciences such as medicine have their own more specialised schemes, for example SNOMED (Systematized Nomenclature of Medicine) is a tree structured taxonomy which organises medical concepts into eleven top level modules such as morphology, diseases, and diagnoses.

3.1.3 Syntax, Semantics and Language

Ontology also plays a role in providing the semantic foundations of language. There have been a number of attempts at producing universal languages and generative ontological knowledge systems. Ramon Llull (ca. 1233–1316, also referred to as Raymundus Lullus) invented a simple mechanism (published in his *Ars generalis ultima* of 1305) consisting of up to 14 concentric disks inscribed by words and symbols. His aim was that by rotating the disks and reading off the words, all (Biblical) truths could be generated mechanically [253]. This and other early attempts inspired Renaissance scientists such as Gottfried Leibniz, whose *Dissertatio*

[2] For example, *Homo sapiens* to refer to modern humans, the only surviving species of the genus *Homo*.



Chap. I. *The General Scheme.* 23

All kinds of things and notions, to which names are to be aſſigned, may be diſtributed into ſuch as are either more

General; namely thoſe Univerſal notions, whether belonging more properly to
Things; called TRANSCENDENTAL
GENERAL. I
RELATION MIXED. II
RELATION OF ACTION. III
Words; DISCOURSE. IV
Special; denoting either
CREATOR. V
Creature; namely ſuch things as were either *created* or *concreated* by God, not excluding ſeveral of thoſe notions, which are framed by the minds of men, conſidered either
Collectively; WORLD. VI
Diſtributively; according to the ſeveral kinds of Beings. whether ſuch as do belong to
Subſtance;
Inanimate; ELEMENT. VII
Animate; conſidered according to their ſeveral
Species; whether
Vegetative
Imperfect; as *Minerals*,
STONE. VIII
METAL. IX
Perfect; as *Plant*,
HERB conſid. accord. to the
LEAF. X
FLOWER. XI
SEED-VESSEL. XII
SHRUB. XIII
TREE. XIV
Senſitive;
EXANGUIOUS. XV
Sanguineous;
FISH. XVI
BIRD. XVII
BEAST. XVIII
Parts;
PECULIAR. XIX
GENERAL. XX
Accident;
Quantity;
MAGNITUDE. XXI
SPACE. XXII
MEASURE. XXIII
Quality; whether
NATURAL POWER. XXIV
HABIT. XXV
MANNERS. XXVI
SENSIBLE QUALITY. XXVII
SICKNESS. XXVIII
Action
SPIRITUAL. XXIX
CORPOREAL. XXX
MOTION. XXXI
OPERATION. XXXII
Relation; whether more
Private.
OECONOMICAL. XXXIII
POSSESSIONS. XXXIV
PROVISIONS. XXXV
Publick.
CIVIL. XXXVI.
JUDICIAL. XXXVII
MILITARY. XXXVIII
NAVAL. XXXIX
ECCLESIASTICAL. XL.

Figure 3.1: Page from the 1668 book *An Essay Towards a Real Character and Philosophical Language* **by John Wilkins.**

de Arte Combinatoria (1666) aimed to reduce all reasoning and discovery to a combination of basic elements. His (unfinished) attempt at producing an "alphabet of human thought" sought to produce a language founded on logic which would provide a sound foundation for the "universal lexicon" which Rene Descartes and others had tried to define. He succeeded in outlining a system called "geometrical characteristic", a calculus for producing concepts of Euclidean geometry.

John Wilkins (1614–1672), a founder of the Royal Society[3], published a book entitled *An Essay Towards a Real Character and Philosophical Language* (London, 1668), in which he introduced a new shorthand orthography (his "real character") for English and sought to provide a means of expressing all known concepts via systematic composition from a set of 40 basic concepts (the *philosophical language*). Based on this universal taxonomy (see figure 3.1), his language was designed such that only ontologically valid sentences could be generated in the language, i.e. that no "false" statement (relative to his ontology) could be made without violating the syntax of the language. This remarkable endeavour was intended to overcome the vagaries of Latin, which was at the time the dominant medium of learned discourse.

In modern computer science parlance, the strings generated by Wilkins' language are thus highly "impure" in the sense that they are designed to uniquely reveal information about the denoted entity or concept rather than being mere arbitrary labels. For example, Wilkins' term for "salmon" is composed from elements of the philosophical language in such a way that the word itself describes the species as "scaled river fish with reddish meat" [28]. Wilkins' language is an early example of an artificial generative descriptive system, since it provides a means of constructing valid descriptions of all possible classes and entities (with respect to the underlying ontology) by repeated application of the syntactic rules. Those statements that violate the rules of the language are by definition invalid also in a semantic sense, i.e. they describe entities which do not or could not exist in reality as defined by the ontology.

However, Wilkins' ontology has many limitations, and many of the distinctions he made seem arbitrary or unjustified by modern standards. Important criteria in

[3] Wilkins also made important contributions to cryptography (a term he introduced into English) and communications. He served briefly as Master of Trinity College, from his appointment by the younger Cromwell in 1659 until the Restoration in 1660.

the design and evaluation of languages and ontologies are the scope and accuracy of the underlying taxonomy and the degree to which statements generated by the system are accurate descriptions of the chosen domain. While this is extremely difficult in the most general case of knowledge in the natural sciences or beyond (such as mathematics or metaphysics), it is possible to design effective ontologies for restricted domains into which we have better cognitive penetrance. The divisions along which any particular ontology is constructed are ill-defined unless some quantifiable criteria are applied to evaluate them, since entities may be grouped and differentiated according to different criteria for different purposes. Borges [28] gives the example of a fictitious Chinese encyclopaedia in which animals are classed into groups such as "those that belong to the Emperor", "stray dogs", "those drawn with a very fine camelhair brush", and "others".

The notion of a formal, unambiguous, and universal language has also been of interest in the field of natural language processing. Such a language could serve as a useful intermediary (known as an "interlingua") for machine translation and automated text understanding. Some artificial human languages such as Esperanto are indeed easier to analyse computationally, but the ability of people to grasp the many subtleties and ambiguities of natural languages through contextual inference and prior knowledge remains unmatched despite very considerable research efforts.

The (now defunct) Cambridge Language Research Unit (CLRU) conducted some influential early research into natural language processing, machine learning, and information retrieval. One of its contributions to machine translation in the 1960s was the definition of a simple formal intermediary language, inspired by the works of Wilkins and others, based on about 80 semantic primitives with simple syntactic relationships between them.

More recently, a number of large scale projects have resulted in systems which continue this type of research. WordNet [175] is a lexical database which establishes semantic relationships such as synonyms, antonyms, hyperyms, hyponyms, and meronyms between about 140000 English words. It functions as a dictionary and thesaurus in which word meaning is described through short natural language descriptions and through relationships (such as synonymy, part-of, subtype-of etc.) with other words. The FrameNet project [81] also provides a lexical database but uses several thousand *semantic frames* [162] to link words and word senses which constitute particular concepts, with frames being related via inheritance. While systems such as WordNet and FrameNet provide some measure of lexical semantics, there are still problems such as completeness and word sense disambiguation.

The even more ambitious CyC project [156] attempts to capture a very large amount of "common sense" knowledge[4] through manually specified assertions expressed in a form of predicate calculus. In order to avoid contradictory inferences, the several thousand concepts represented by CyC are grouped into thousands of "microtheories", which are internally consistent specifications of a particular knowledge domain arranged in specification hierarchies (e.g. geometry being a specialised subdomain of mathematics). Though more than one million such rules have been defined over the past 20 years, the project has been criticised on the grounds of scalability, consistency, and uncertainty handling problems.

While the sheer scale of these systems makes them useful for many problems in knowledge modelling and language analysis, they ultimately provide recursive definitions of concepts in terms of other concepts with no ability to link these to the world via sensory inputs. Although lexical disambiguation of text can be performed on the basis of statistics and can benefit from the knowledge these systems provide [122], they do not by themselves provide a way of handling non-textual information such as that processed in computer vision. By contrast, the ontological languages and inference methods proposed in this dissertation encode knowledge about a domain both *intensionally*, i.e. through syntactic rules and relationships between terms, and *extensionally* in terms of the processing modules required to infer or recognise the denoted concepts, entities, and properties on the basis of visual and other sources of information.

3.1.4 Ontologies and Languages in Computing

In the computer and information sciences, ontologies often take the form of schemata that characterise a given domain. These are generally far more specialised than those mentioned above in that the data structures and relationships encoded in a particular program or database are usually custom built for a particular problem

[4] Examples include "animals live for a single period of time", "nothing can be in two places at once", "every tree is a plant".

or application, rather than attempting to capture abstract or scientific notions of truth.

Terms such as data, information, and knowledge are often loosely defined and sometimes used interchangeably, although attempts have been made to arrive at a more precise definition [3, 38]. Data has no meaning or structure (other than type) on its own and can be regarded as consisting of symbols from an alphabet or numerical representations of some quantity. Information can be seen as patterns of data (e.g. as a structured message or as a non-arbitrary transformation of some signal), which give it the syntactic and statistical qualities through which it is quantified in information theory and systems theory. Knowledge is often regarded as requiring information put to some purpose such as the interpretation of information to ascribe semantics, the solution of some problem, or the formation of new knowledge. Etymologically, the word data is the plural of the Latin *datum*, meaning “something given”, whereas knowledge or insight must be derived in some way.

Ontologies may therefore be regarded as repositories of information linking together particular facts or through which data can be given a particular interpretation, i.e. knowledge arises in a particular context. The knowledge content of an ontology may be *factual*, i.e. encoded in the ontology itself, or *inferential*, which means that it must be derived by some form of reasoning. A further distinction can be made between *procedural* and *declarative* means of implementing ontologies in computer systems [248]. The former is built upon the fundamental computer science concept of *process* [38], and seeks to provide computer programs with active “know-how” in the form of specialised software components for the various parts of the ontology. The latter focuses on representations of “know-that” in the form of data structures or databases and logical axioms.

Ontologies can thus be seen as a widely used tool in computer science as a whole, although their usage is made most explicit in the field of artificial intelligence (AI) and related disciplines such as robotics and knowledge engineering (KE) [278]. Many different formalisms exist for expressing and utilising ontologies [232, 258]. Object-oriented programming techniques [262] have become commonplace. Their representation methodology in terms of classes, class attributes, objects (which are instantiations of classes), and class inheritance hierarchies has influenced (and been influenced by) a number of languages and schemes used for digital knowledge representation. Many database systems are turning from the classical relational model (based on predicate logic and set theory) to object-relational and object database schemes.

3.1.5 Knowledge Representation and Reasoning

In the AI and KE domains, several formalisms for knowledge representation and management have become popular [106, 232]. A *semantic network* is a directed graph consisting of vertices that represent concepts and edges that encode semantic relations between them. Concepts can be arranged into taxonomic hierarchies and have associated properties. *Frame systems* are closely related to semantic networks but represent knowledge in terms of hierarchies of frames containing named slots, together with rules such as type constraints, to define concepts and relationships. Much recent work on knowledge modelling has focused on *description logics*, which have evolved from semantic networks to provide a more rigorous definition of semantics in terms of a (typically restricted) form of first-order logic. Description logics usually provide a syntax that makes it easy to specify categories and perform inference tasks such as subsumption and classification.

Logics of various kinds and logical reasoning and representation languages such as Prolog and KL-ONE have been popular tools for knowledge modelling, for example in the definition of *expert systems*. However, the problems of linking and encapsulating real world concepts by means of logic and related formalism led to widespread disappointment in the gap between early claims of AI enthusiasts and the actual performance of many systems.

Apart from the difficulties [195, 100] of acquiring, formulating, and translating between different representations of concepts, two important problems in this respect are the *frame problem* [170, 63] and the *symbol grounding problem* [104]. The frame problem originally concerned the difficulty of using logic to represent which facts of the environment change and which remain constant over time. The term is now frequently used in a broader scope to refer to the need to limit the number of assertions and inferences that an artificial agent must make to solve a given task in a particular context. The symbol grounding problem concerns the difficulty of relating abstract symbols to aspects of the real world. Much of the earlier work in

AI assumed that intelligence could be achieved by abstract reasoning over symbols, however more recent work increasingly recognises that the categories and objects which people take for granted are extremely difficult to formalise satisfactorily [38], and that linking them to concrete sensory data is a considerable challenge for any perceiving system [41, 12]. The fact that many AI systems incorporate within them "microtheories" that are not clearly or formally linked to external reality also makes it very difficult to integrate such systems since they do not share a common view of the world.

While efforts in AI have often relied on formalisms such as mathematical logic as a means of representing knowledge *intensionally*, much of the recent work in machine learning [280, 82] can be seen as providing *extensional* definitions of concepts in terms of classifiers and recognisers. Given a training corpus, these methods provide a means of deciding (with a certain probability of success) whether a given example is an instance of a particular concept or not. They do not generally produce an explanation of why it is a member of that category and do not incorporate a description of the properties of the object. In fact many statistical learning paradigms such as support vector machines rely on simplifying and abstracting the classification problem to minimise the amount of information about the problem that needs to be represented while maximising the separability of the different classes. A purely intensional representation on the other hand consists of rules or statements whose truth may be assessed independently of the particular extensions involved, i.e. regardless of which element of the modelled class is being considered. There are some attempts at combining these two approaches, for example *formal concept analysis* [50, 73] that has been applied successfully to some text categorisation and retrieval problems. The work presented in this thesis can be seen as providing a means of representing and matching knowledge in computer vision and other domains both intensionally and extensionally.

3.1.6 The Semantic Web and Beyond

The term *ontology* has recently undergone a strong revival largely due to the efforts of the *semantic web*[5] community. Ontologies are seen in a knowledge management

[5] See *http://www.w3.org/2001/sw/* and *http://www.semanticweb.org/*.

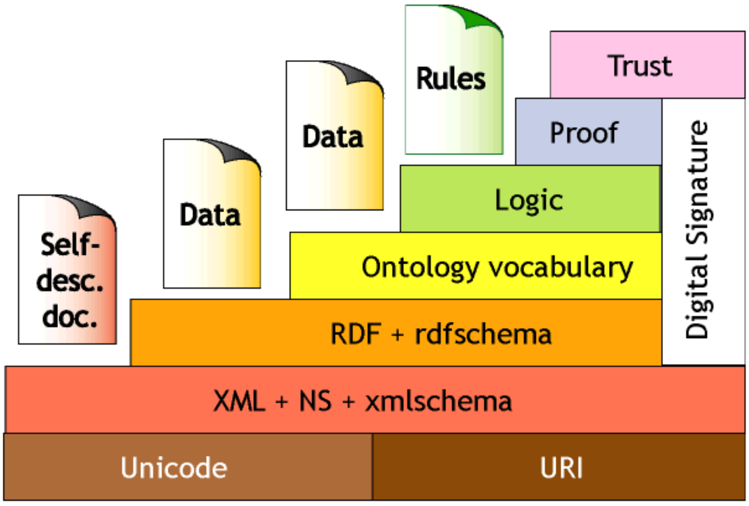


Figure 3.2: Layered content architecture for the *Semantic Web*.

sense as providing an important tool for the representation of semantic information and the automated processing of such information for applications such as data mining, retrieval, and automated discovery and utilisation of services by autonomous software agents. The consensus definition of ontology in this context is as a "formal, explicit specification of a shared conceptualisation" [97], hence the focus is on knowledge sharing and sufficiently formal representation to allow manipulation by computer.

As shown in figure 3.2, the semantic web is based on a layered hierarchy in which ontologies provide the underpinning that enables metadata to be interpreted automatically [80, 79, 168, 258]. In addition to the XML and RDF standards for structured document annotation and resource description, attention is now focussed on the new framework for Web Ontology languages (OWL)[6]. It is now recognised that ontologies are the natural vehicle for knowledge representation and interchange at different levels of granularity and across different domains, hence they are to form a vital cornerstone of future generations of the internet.

While the proposed formalisms for the semantic web draw on a rich heritage of work in artificial intelligence and linguistics, they remain limited due to an explicit focus on the description of textual document resources. Many decades of research

[6] http://www.w3.org/2001/sw/WebOnt/.

into knowledge engineering have shown the limitations and brittleness of methodologies such as semantic nets and description logics on which OWL is based. The World Wide Web Consortium (W3C) and the knowledge engineering community have yet to address the enormous challenge of reasoning about non-textual information domains such as the vast quantities of video and image data now prevalent across the internet and private networks. Annotation schemas expressed in terms of XML and RDF are inherently incomplete, ambiguous, and non-extensible due to the fact that they are no longer grounded in the data which they seek to describe.

By contrast, representations expressed in the ontological languages proposed in this thesis retain a linkage to their domain of discourse through explicit grounding of terminal symbols in terms of extracted image features, labelled image and video regions, or SQL statements that enable relevant content to be referenced dynamically. The abstract syntax tree representation allows relationships of arbitrary arity and level of abstraction to be defined. Higher-level semantics are represented through Bayesian networks, which provide a powerful framework for inference and reasoning under uncertainty that largely avoids the pitfalls of the deterministic first-order logic formalism proposed for OWL.

3.2 Proposed Approach and Methodology

3.2.1 Overview

This dissertation proposes an ontology-based architectural model for high-level vision. 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* [104], 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* [63], 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 that 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 that 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.

3.2.2 High-level Vision

While robust high-level interpretation of visual information remains a very challenging and ill-defined endeavour, greater progress has been made in deriving low-level features from pixels and arriving at region-based image representations. Regions are defined as connected parts of an image which share a common set of properties, and the term segmentation refers to the process of finding a covering set of non-overlapping regions based on image features that are usually defined in terms of colour, shape, and texture properties. The goal is to find regions such that there is greater variation between neighbouring regions than within individual ones. The next logical step in narrowing the “semantic gap” between computer representations and human perception of images lies in the definition of objects corresponding to semantically meaningful entities in a scene. A distinction must here be made between region segmentation and object segmentation, since each object may comprise a number of regions and each region may contain several objects.

While the region segmentation of an image will be relatively well-defined for a particular choice of feature predicates and segmentation algorithm, the problem of finding a suitable object-labelling is much less well understood and depends greatly on the prior assumptions one makes about what constitutes an object. Purely low-level features such as edges are generally not enough for object segmentation, since object boundaries are often recognised by human observers based on prior knowledge of object-level compositional semantics and the integration of contextual cues.

In the case of video data, one also has to deal with variation over time, which may be caused by motion (of the camera, the objects in the scene, or both) but may also be the result of differences in lighting or the motion of objects in other parts of the scene, possibly outside the camera’s viewpoint (the “aperture problem”). In order to track regions or objects across multiple frames, one needs to establish a correspondence between features in temporally displaced images. Issues such as occlusions, shading, background motion, motion discontinuities, and the sheer volume of data exacerbate the problem. However, image sequences also offer additional continuity constraints that can prove beneficial for content-extraction and analysis tasks.

As illustrated in figure 3.3, the content-based representation of digital images by a computer can be viewed in terms of a hierarchical arrangement corresponding to increasingly higher levels of abstraction oriented towards capturing its semantic interpretation, i.e. the “meaning” of the image and the objects that are represented by it. In the case of video sequences such an interpretation naturally encompasses

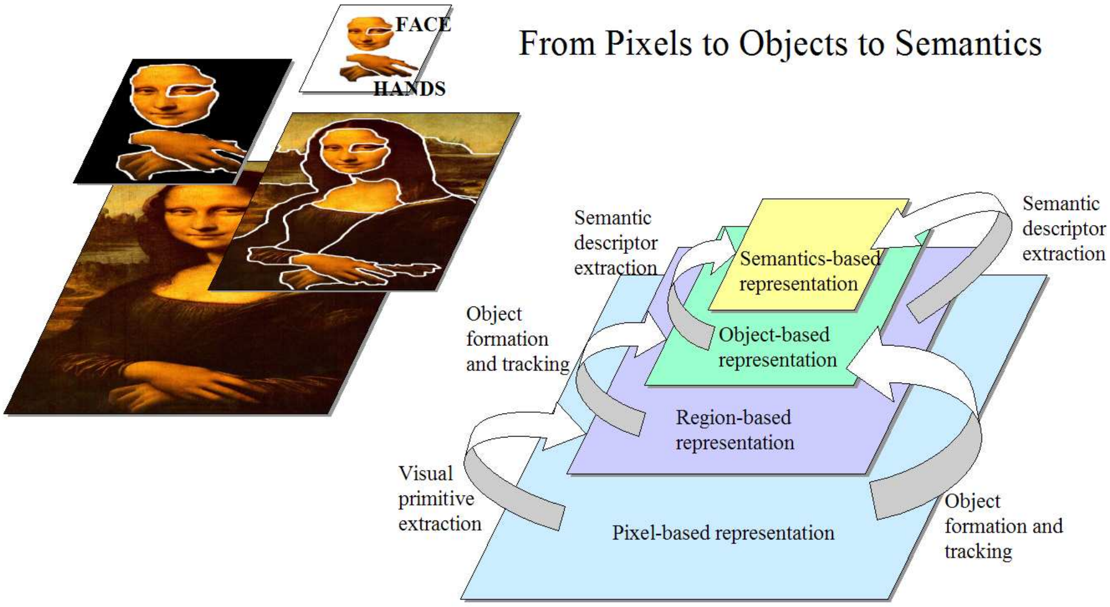

FACE
HANDS
From Pixels to Objects to Semantics
Semantic descriptor extraction
Semantics-based representation
Semantic descriptor extraction
Object formation and tracking
Object-based representation
Region-based representation
Visual primitive extraction
Pixel-based representation
Object formation and tracking

Figure 3.3: Hierarchical representation of visual information at different conceptual levels and example illustrating these concepts as applied to an image of the "Mona Lisa" (original copyright: Musee de Louvre).

motions, actions, interactions, and composite events. This "bottom-up" approach to vision generally proceeds by identifying features, using such features to group pixels into regions, recognising which regions contain or are part of objects, and finally ascribing some semantics to the scene or sequence. Conversely, this process can also be carried out "top-down", i.e. starting with some semantics to constrain the domain of interest, and using such constraints to search for objects and features.

Vision is clearly a challenging task, made difficult due to the great variability of visual information and problems such as ambiguity, noise, and complexity. A number of broad schools of thought exist regarding the nature of the central problems in vision and how these may be solved in biological and artificial vision systems. A systematic description of these is clearly beyond the scope of this thesis, however the following sketches some of the key ideas (see chapter 2 for further references):

- *Vision as Finding What is Where*: Vision is the primary sensory modality of many animals including humans. It is what people predominantly rely on when they see, look at, perceive, behold, admire, glance at, view, eye, observe, glimpse at, make out, witness, spot, or sight something. One of its core functions is clearly the rapid and reliable identification and localisation of pertinent features of the environment such as predators, food, and other animals. The human brain appears to rely heavily on feedback mechanisms to solve these tasks and provides a separate processing path for "what" (via the inferotemporal cortex) and "where" (via the parietal cortex) [96].

- *Vision as Inverse Graphics*: In a sense vision is the inverse problem of graphics, i.e. it entails the challenge of inferring objects and attributes from a particular rendering of a scene. As such, vision is an ill-posed and generally heavily underdetermined problem that requires a variety of constraints to define a plausible solution in a particular instance. This view inspired many of the early approaches to computer vision, for example Marr's formulation of the problem in terms of a mapping from images to 3D geometric objects via feature detection and the "2.5 dimensional sketch" [167]. Vision appears to be an "AI complete" problem, meaning that solving it would entail solving all the great problems of artificial intelligence such as language understanding, learning, reasoning, and planning.

- *Vision as Graphics*: However, there is also a sense in which vision is like graphics, an active generative process by which perceiving systems construct an internal model of the world based on their perceptions, prior knowledge, and current goals. Optical illusions are a rich source of evidence for this, as are more recent studies of how the brain processes visual input [96]. Early scientists of vision such as Hermann von Helmholtz (1881) noted that "The meaning we assign to our sensations depends upon experiments and not upon mere observation of what takes place around us." Vision is therefore an active exploratory process rather than a passive filtering and transformatory activity.

- *Vision as Inference*: In all of the above formulations, vision requires one to infer properties of the world based on visual input and prior knowledge.

As noted previously, probabilistic inference techniques such as Bayesian networks provide a principled framework for handling uncertainty and complexity in artificial perceiving systems. There are many debates and open questions regarding the kind of representations that best support such inferences, be they object or viewer centred, appearance-based or model-based, or a combination of these. The crucial role of learning has also gained prominence in artificial vision research, and is closely related to issues of representation.

- *Vision as Language Processing*: As mentioned above, many problems in vision such as object recognition [70], video analysis [113, 209, 148], gesture recognition [25, 130, 179], and multimedia retrieval [133, 299, 16, 270] can be viewed as relating symbolic terms to visual information by utilising syntactic and semantic structure in a manner related to approaches in speech and language processing [2, 257, 227]. The importance of language in shaping the way humans process visual and other information has also been the subject of research in the brain and cognitive sciences [12, 13, 41].

A visual language can also serve as an important mechanism for attentional control by constraining the range of plausible feature configurations which 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 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 that are of value given a particular query are evaluated, and evaluation may stop as soon as the appropriate top level symbol sequence has been generated.

3.2.3 Recognition and Classification

The notion of image and video interpretation relative to the goal of satisfying a structured user query (which may be explicit or implicitly derived from a more general specification of system objectives) follows the trend in recent approaches to robust object recognition on the basis of a "union of weak classifiers". Such an approach hierarchically integrates trained parts-based relationships between lower level feature classifiers to recognise composite objects.

This idea has a rich heritage in the field of Gestalt psychology [292], founded by Köhler, Wertheimer, and Koffka in the 1920s, which describes how image structure is perceived based on principles such as proximity, similarity, continuity, and coherence. Much recent work in computer vision can be seen as incorporating such ideas, particularly as regards the perceptual organisation of image content (e.g. [78], [83], [137], [235]). Perceptual grouping in this context is the process of extracting salient structural image information by clustering features through the effective use of multiple cues, the combination of local and global measurements, and the incorporation of domain specific prior knowledge. Most efforts to find perceptual groupings of image features focus on finding *non-accidental* image structure as identified by means of a particular set of predicates over lower-level image properties (e.g. texture, shape, colour). Making such methods robust, scalable, and generally applicable has proven a major problem. In Koffka's [147] words:

> *"...to apply the Gestalt category means to find out which parts of nature belong as parts to functional wholes, to discover their position in these wholes, their degree of relative independence, and the articulation of larger wholes into sub-wholes."*

However, there is a principle of Gestalt theory that is less frequently mentioned in computer vision research. It is the notion that objects are not exhaustively describable solely on the basis of their constituent parts and that any attempt at recognising objects based on pre-defined categories or feature sets is inherently limited in its applicability. Max Wertheimer [292] captures this idea by stating:

> *"There are wholes, the behaviour of which is not determined by that of their individual elements, but where the part-processes are themselves*

> *determined by the intrinsic nature of the whole (...) This problem cannot be solved by listing possibilities for systematisation, classification, and arrangement. If it is to be attacked at all, we must be guided by the spirit of the new method and by the concrete nature of the things themselves which we are studying, and set ourselves to penetrate to that which is really given by nature."*

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 thesis 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 determining non-accidental and perceptually significant visual properties necessarily depend on the objectives and prior knowledge of the observer, as recognised in [26]. 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.

3.2.4 Proposed Framework

In spite of the benefits of Bayesian networks and related formalisms outlined above, probabilistic graphical models also have limitations in terms of their ability to represent structured data at a more symbolic level [207, 206] 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

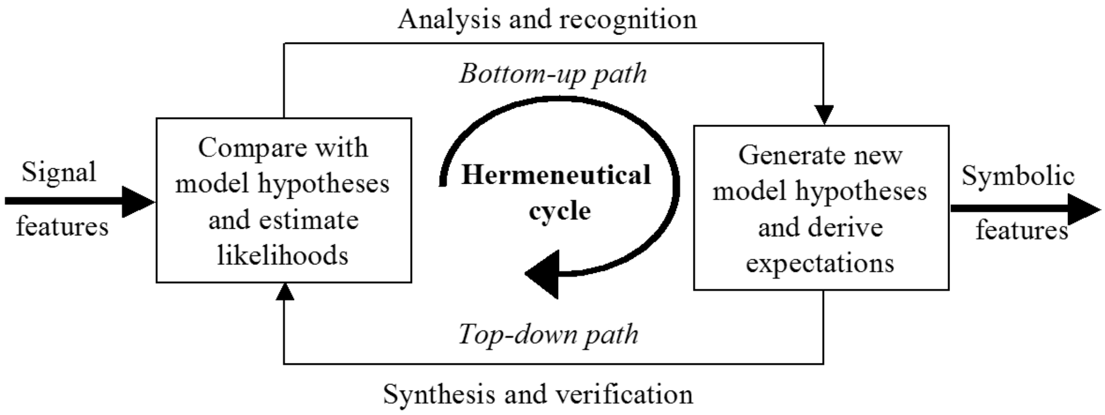

Analysis and recognition
Bottom-up path
Signal
features
Compare with model hypotheses and estimate likelihoods
Hermeneutical cycle
Generate new model hypotheses and derive expectations
Symbolic
features
Top-down path
Synthesis and verification

Figure 3.4: The Hermeneutical cycle for iterative interpretation in a generative (hypothesise and test) framework.

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 3.4 illustrates this approach.

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 [35]. 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 thesis 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

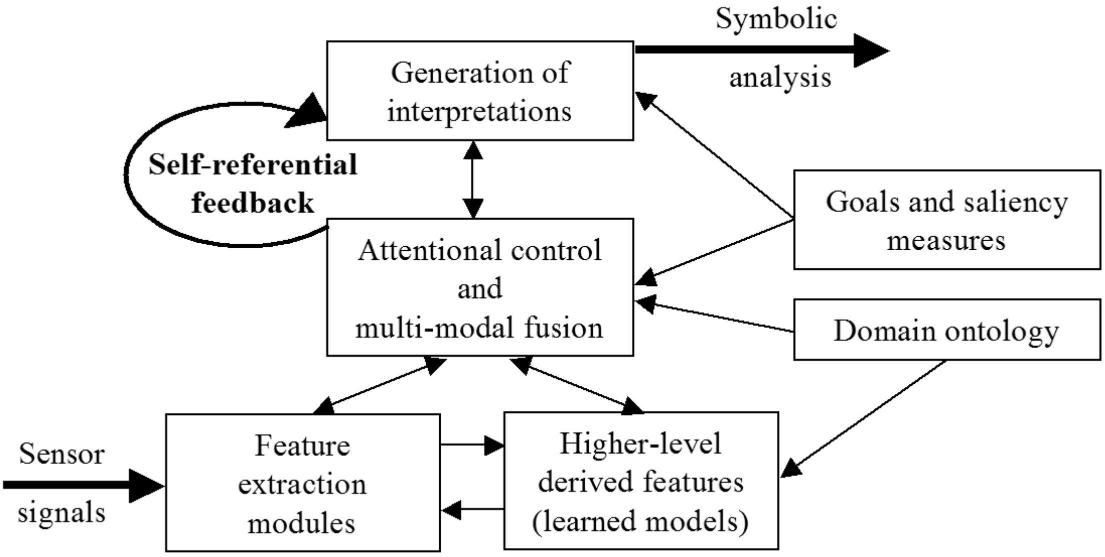

Symbolic
analysis
Generation of interpretations
Self-referential feedback
Goals and saliency measures
Attentional control and multi-modal fusion
Domain ontology
Sensor
signals
Feature extraction modules
Higher-level derived features (learned models)

Figure 3.5: 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.

3.5, 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.

Chapter 4

Ontological Query Language for Content-based Image Retrieval

4.1 Overview

This chapter 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 that 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.

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 means of addressing the key problems of image retrieval, namely the ambiguity of image content and user intentions, as well as the semantic gap that exists between user and system notions of relevance. It represents a radical departure from existing image query interfaces such as those based on sketches, example images, feature predicates, annotations, or document context.

4.2 Introduction

Advances in technology such as digital cameras, scanners, and storage media have led to a proliferation of digital image collections, and increased internet access and bandwidth make a vast quantity of such data available to a wide audience. However, many such collections lack effective indexing or annotation schemes and it is consequently very difficult to organise, browse, or retrieve images based on their visual content. Image data continues to grow at a rate which greatly outpaces that of text documents, yet there is a lack of viable means of searching and organising such data to leverage its value. As in the days before the computer age, when "computers" were actually people hired to perform calculations, an "image searcher" today is typically a human being who often needs to browse through thousands of images in the hope of finding what he or she is looking for. The group of people performing such searches ranges from professional image archivists working for commercial image collections, newspapers, museums etc., to scientific researchers, medical practitioners, lawyers, police and intelligence officers, photographers, and home users.

Despite great advances in the fields of information retrieval, particularly of text documents (see [255, 220]), and computer vision in recent years, there are few if any truly robust general purpose methods that offer a solution to the challenges of content-based image retrieval. Image retrieval suffers from additional complexity and uncertainty arising from the fact that the salient conceptual features represented in a two-dimensional image are fundamentally underdetermined [254]. One therefore has to deal with a far greater degree of ambiguity than is the case with natural language terms, and problems such as feature extraction and labelling, object recognition, etc. present enormous difficulties. Moreover, issues such as request formulation, query refinement, relevance feedback, and performance evaluation are far less tractable and not as well understood as is the case with text retrieval.

Most research systems for content-based image retrieval (CBIR) generally only provide search over very abstract or low-level image features such as colour distributions or statistical shape measures. Furthermore, they require users to supply an example image of the exact type they are searching for or employ other cumbersome query interfaces based on sketching or selection of feature thresholds.

Current commercial image retrieval systems therefore largely rely on manual annotations or other context such as captions to reduce the problem to one of text-based retrieval over image descriptions. However, generating such descriptions is very time-consuming and fraught with problems due to human error, differing interpretations, and the intrinsic ambiguity and immutability of written language. Attempts at automated image description fail to capture the intuitive understanding of human observers.

Query mechanisms play a vital role in bridging the *semantic gap* [99] between users and retrieval systems. There has however been relatively little recent work in addressing this issue in the context of content-based image retrieval (CBIR). Most of the query interfaces implemented in current systems fall into a small group of approaches. In order to overcome the weaknesses of these methodologies, efforts have focused on techniques such as *relevance feedback* [293, 53, 234, 109] as a means of improving the composition and performance of a user query in light of an initial assessment of retrieval results. While this approach and other methods for improving the utility of user queries by means such as query expansion and through the combination of multiple query modalities have shown some promise, they do so at the risk of increased user effort and lack of transparency in the retrieval process.

This chapter presents the notion of an ontological query language as a powerful and flexible means of providing an integrated query and retrieval framework that addresses the problem of the *semantic gap* between user and system. Further background information and a general motivation for such query languages is given in section 4.3, while section 4.5 introduces the *OQUEL* language as a concrete example for retrieval from photographic image collections. Section 4.4 discusses the basic language design and structure. In section 4.6 the process of query interpretation and retrieval is described further. The discussion is based on an implementation of the language for the *ICON*[1](Image Content Organisation and Navigation, [269]) content-based image retrieval system. Those content extraction and representation facilities of ICON relevant to the present discussion are outlined in sections 4.5.2 to 4.5.6. Section 4.7 gives quantitative performance

[1] Parts of ICON were originally written by the author of this thesis during the course of his employment at AT&T Laboratories.

results of OQUEL queries compared to other query modalities in the ICON system. The chapter concludes with a discussion of results and summary in section 4.8, which also provides an outlook of further developments.

4.2.1 CBIR Query Mechanisms and the Retrieval Process

As has been noted elsewhere (e.g. [247]), research in content-based image retrieval has in the past suffered from too much emphasis being placed on a system view of the retrieval process in terms of image processing, feature extraction, content representation, data storage, matching, etc.. It has proven fruitful in the design of image retrieval systems to also consider the view of a user confronted with the task of expressing his or her retrieval requirements in order to get the desired results with the least amount of effort. While issues such as the visualisation of query results and facilities for relevance feedback and refinement are clearly important, this section is primarily concerned with the mechanisms through which users express their queries.

Adopting a user perspective, one can summarise most of the query methods traditionally employed by CBIR systems (see [247] and [231] for further references) and highlight their drawbacks as follows:

- *Query-by-example*: ([265], [53], [145]) Finding suitable example images can be a challenge and may require the user to manually search for such images before being able to query the automated system. Even when the user can provide images that contain instances of the salient visual properties, content, or configurations they would like to search for, it is very hard for the system to ascertain which aspects make a given image relevant and how similarity should be assessed. Many such systems therefore rely on extensive relevance feedback to guide the search towards desirable images, but this approach is not appropriate for most real-world retrieval scenarios. Many industrial applications of CBIR require ways of succinctly expressing abstract requirements which cannot be encapsulated by any particular sample image.

- *Template, region selection, or sketch*: ([44], [143], [39]) Rather than providing whole images, the user can draw (sometimes literally) the system's attention to particular image aspects such as the spatial composition of desired content in terms of particular regions or a set of pre-defined templates. Clearly this process becomes cumbersome for complex queries and there are difficult user interface issues concerning how one might best represent abstract relations and invariants.

- *Feature range or predicate*: ([192], [189]) Here the user can set target ranges or thresholds for certain (typically low-level) attributes such as colour, shape, or texture features that may represent global image properties or features localised to certain image regions. While this clearly has merit for some types of queries, the approach requires a certain amount of user sophistication and patience and is ill-suited to retrieval based on higher-level concepts.

- *Annotation or document context*: ([176], [250], [120]) Images rarely come with usable annotations for reasons such as cost, ambiguity, inconsistency, and human error. While progress has been made in applying text retrieval methods to annotations and other sources of image context such as captions, difficulties remain due to lack of availability, unreliability, and variability of such textual information. Image annotations cannot be changed easily to reflect the particular point of view mandated by a given query, nor are they an effective means of representing visual properties. There is scope for methods which combine image descriptions for concepts that are almost impossible to assess by vision techniques alone (e.g. artist, period, and “mood” of a painting) with features extracted by computer vision [16, 70, 15].

- *Query language or concept*: ([49], [186], [234]) Efforts have been made to extend popular database query languages derived from SQL to cater for the intrinsic uncertainty involved in matching image features to assess relevance. However, such languages remain quite formal and rigid and are difficult to extend to higher-level concepts. Knowledge-based approaches utilising description logics or semantic networks have been proposed [118, 231] as a means of better representing semantic concepts but tend to entail somewhat cumbersome query interfaces.

Although these approaches have proven to be useful, both in isolation and when combined, in providing usable CBIR solutions for particular application domains and retrieval scenarios, much work remains to be done in providing query mechanisms that will scale and generalise to the applications envisaged for future mainstream content-based access to multimedia. Indeed, one criticism one can generally level at image retrieval systems is the extent to which they require the user to model the notions of content representation and similarity employed by the system, rather than vice versa. One reason for the failure of CBIR to gain widespread adoption is due to the fact that mainstream users are quite unwilling to invest great effort into query composition [221, 222] as many systems fail to perform in accordance with user expectations.

The language-based query framework proposed in this chapter aims to address these challenges. Query sentences are typically short (e.g. "people in centre") yet conceptually rich. This is because they need only represent those aspects of the target image(s) which the user is trying to retrieve and which distinguish such images from others in the dataset. The user is therefore not required to translate a description of an envisaged target image into the language but merely (and crucially) to express desired properties that are to hold for the retrieved images. Hence even a fairly short query sentence can suffice to select a small subset of desired images from a vast collection. This simple idea is the reason why text retrieval on the internet is so successful: the less frequently a particular constellation of keywords appears across the entire document set, the more valuable it is as a means of discriminating relevant from non-relevant content.

These insights give rise to a number of factors that can be used in designing and assessing query mechanisms, including:

- *Ease of expression*: The query interface should allow users to concisely state their retrieval requirements with minimal effort, both conceptual (cost of translation to given query format) and manual (complexity of query composition task) effort. Queries should be as simple as possible and as complex as necessary.

- *Ease of understanding*: Conceptual transparency in the retrieval process is desirable to ensure that users can comprehend the system's response to their queries and formulate queries accordingly. This correspondence is made easier if queries relate to concepts of which users already have an intuitive understanding without giving a false sense of the system's capabilities. A lack of understanding of what facilities the system provides and how queries should be constructed inevitably leads to user frustration.

- *Generality and extensibility*: In order to cater for a wide range of retrieval scenarios and users, the query interface should provide exposure to both the low-level and high-level content description facilities supported by the underlying system. Experience has shown that image retrieval works best on limited specialised domains [89, 60], but an intrinsic flexibility is desirable so that individual users may adapt the system to their own needs and incorporate relevant domain knowledge [49].

- *Responsiveness*: The interface should provide fast user feedback even while a query is being processed. Allowable processing times are task-dependent, i.e. a professional image librarian trying to satisfy a customer request may well tolerate a greater delay than a more casual user engaged in some browsing activity.

- *Context awareness*: Different query and retrieval mechanisms are appropriate depending on the context in which the search takes place. Factors include: *search behaviour* - the differences between casual browsing, broad similarity search, search for a specific target image, search for images meeting given criteria, etc.; *scope* - search in large and unfamiliar (e.g. web, image archive) collections vs. search over a personal image set; *user motivation* - home user, journalist, professional image search (e.g. stock photograph libraries such as Getty Images and Corbis), specialist image analysis task (e.g. medical diagnosis [263] or forensic image search [89]).

There are other issues arising from real-world deployment of retrieval systems such as the ability to integrate a new query mechanism with an existing legacy system, and these will be briefly discussed in section 4.5.8.

4.2.2 Language-based Querying

Query languages constitute an important avenue for further work in developing CBIR query mechanisms. Powerful and easy-to-use textual document retrieval systems have become pervasive and constitute one of the major driving forces behind the internet. Given that so many people are familiar with the use of simple keyword strings and regular expressions to retrieve documents from vast online collections, it seems natural to extend language-based querying to multimedia data. Indeed, there has been a broader trend in recent years of applying techniques familiar from textual document retrieval, such as latent semantic indexing [299], to multimedia information retrieval.

However, it is important to recognise [254] that the natural primitives of document retrieval, words and phrases, carry with them inherently more semantic information and characterise document content in a much more redundant and high-level way than the pixels and simple features found in images. This is why text retrieval has been so successful despite the relative simplicity of using statistical measures to represent content *indicatively* rather than *substantively*. Image retrieval addresses a much more complex and ambiguous challenge, which is why this chapter proposes a query method based on a language that can represent both the *syntax* and *semantics* of image content at different conceptual levels.

This chapter will show that by basing this language on an ontology one can capture both concrete and abstract relationships between salient image properties such as objects in a much more powerful way than with the relatively weak co-occurrence based knowledge representation facilities of classical statistical information retrieval. Since the language is used to express queries rather than describe image content, such relationships can be represented explicitly without prior commitments to a particular interpretation or having to incur the combinatorial explosion of an exhaustive annotation of all the relations that may hold in a given image. Instead, only those image aspects that are of value in determining relevance given a particular query are evaluated and evaluation may stop as soon as an image can be deemed irrelevant.

Content-based image retrieval on the basis of short query sentences is also likely to prove more efficient and intuitive than alternative query composition schemes such as iterative search-by-example and user sketches which are employed by most current systems. Goal-directed natural language-like access to information resources has already found some application in particular domains such as circuit design [203], medical image collections [263], and agricultural data [19]. There is also a relatively long lineage of natural language interfaces to database systems (e.g. [112]), although their success has been rather limited.

The comparatively small number of query languages designed for CBIR have largely failed to attain the standards necessary for general adoption. A major reason for this is the fact that most language or text-based image retrieval systems rely on manual annotations, captions, document context, or pre-generated keywords, which leads to a loss of flexibility through the initial choice of annotation and indexing. Languages mainly concerned with deriving textual descriptions of image content [2] are inappropriate for general purpose retrieval since it is infeasible to generate exhaustive textual representations that contain all the information and levels of detail that might be required to process a given query in light of the user's retrieval need. Recent attempts at solving the inverse generative task of graphically rendering scenes from inherently ambiguous natural language descriptions [54] show promise, but such techniques have yet to be applied to image retrieval.

While keyword indexing of images in terms of descriptors for semantic content remains highly desirable, semi- or fully automated annotation is currently based on image document context [240] or limited to low-level descriptors. More ambitious "user-in-the-loop" annotation systems still require a substantial amount of manual effort to derive meaningful annotations [291]. Formal query languages such as extensions of SQL [224] are limited in their expressive power and extensibility and require a certain level of user experience and sophistication.

In order to address the challenges mentioned above while keeping user search overheads to a minimum, this chapter presents the *OQUEL* query description language. It provides an extensible language framework based on a formally specified grammar and an extensible vocabulary that are derived from a general ontology of image content in terms of categories, objects, attributes, and relations. Words in the language represent predicates on image features and target content at different semantic levels and serve as nouns, adjectives, and prepositions. Sentences are prescriptions of desired characteristics that are to hold for relevant retrieved images. They can represent spatial, object compositional, and more abstract relationships between terms and sub-sentences. The language therefore differs in a number of respects from related attempts at using language or semantic graphs to facilitate content-based access to image collections ([43], [204], [172], [49], [48]).

The language is portable to other image content representation systems in that the lower level words and the evaluation functions which act on them can be changed or re-implemented with little or no impact on the conceptually higher language elements. It is also extensible since new terms can be defined both on the basis of existing constructs and based on new sources of image knowledge and metadata. This enables definition of customised ontologies of objects and abstract relations. The process of assessing image relevance can be made dynamic in the sense that the way in which elements of a query are evaluated depends on the query as a whole (information flows both up and down) and any domain specific information with respect to the ontological makeup of the query which may be available at the time it is processed.

4.3 Ontological Language Framework

4.3.1 Role of Ontologies

By basing a retrieval language on an ontology, one can explicitly encode ontological commitments about the domain of interest in terms of categories, objects, attributes, and relations. Gruber [97] defines the term ontology in a knowledge sharing context as a “formal, explicit specification of a shared conceptualisation”. Ontologies encode the relational structure of concepts that one can use to describe and reason about aspects of the world.

Sentences in a language built by means of an ontology can be regarded as active representational constructs of information as knowledge. There have been similar approaches in the past applying knowledge-based techniques such as description logics [101, 10, 18] to CBIR. However, in many such cases the knowledge-based relational constructs are simply translated into equivalent database query statements such as SQL [118], or a potentially expensive software agent methodology is employed for the retrieval process [66]. This mapping of ontological structures onto real-world evidence can be implemented in a variety of ways. Common approaches are heavily influenced by methods such as description logics, frame-based systems, and Bayesian inference [80].

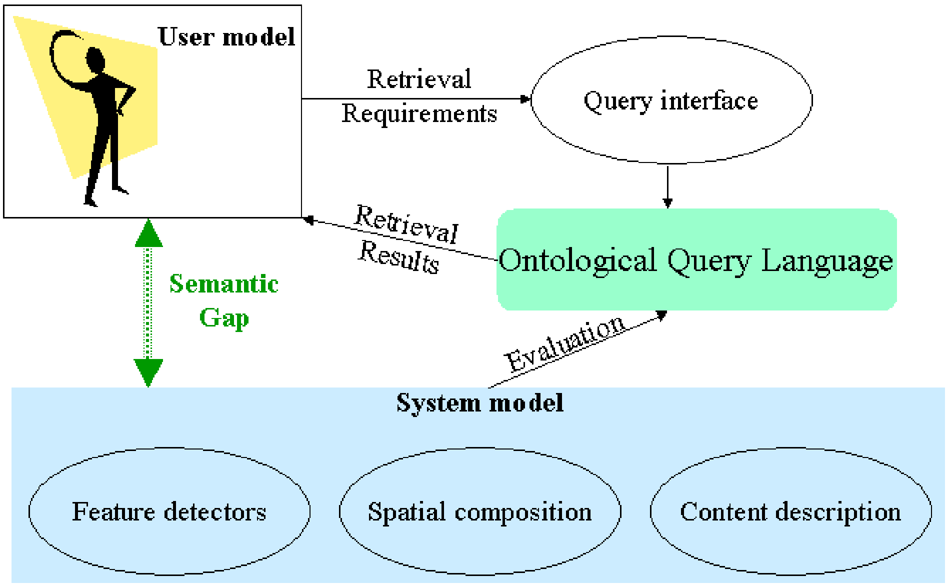


Figure 4.1: Model of the retrieval process using an ontological query language to bridge the semantic gap between user and system notions of content and similarity.

This chapter argues that the role of a query language for CBIR should be primarily *prescriptive*, i.e. a sentence is regarded as a description of a user's retrieval requirements that cannot easily be mapped onto the description of image content available to the system. While the language presented here is designed from a general ontology which determines its lexical and syntactic elements to represent objects, attributes, and relations, this does not in itself constitute a commitment to a particular scheme for determining the semantic interpretation of any given query sentence. The evaluation of queries will depend on the makeup of the query itself, the indexing information available for each image, and the overall retrieval context. Evaluation therefore takes place within a particular ontological domain specified by the composition of the query and the available image evidence at the time it is processed. This approach is consistent with the view expressed in e.g. [234] that the *meaning* of an image is an emergent property that also depends on both the query and the image set over which the query is posed. Figure 4.1 shows how the ontological query language and the mechanisms for its interpretation can thus be regarded as acting as an intermediary between user and retrieval system in order to reduce the semantic gap.

Furthermore, the representation of a general ontology and a particular instance (as defined by the query) of it are independent of the actual implementation of its constituent parts in terms of data structures and matching modules. Hence such a query language can be ported to other CBIR systems or applied to highly heterogeneous image sets without changing the language structure.

The notion of an ontological query language is therefore well suited to addressing the issues outlined in section 4.2.1. Other taxonomies of the query specification process are possible, e.g. [247] groups queries into six categories depending on whether they are exact (only images meeting the specified criteria are retrieved) or approximate (retrieved images are ranked according to some degree of match) and then depending on whether they relate to spatial image content, global image properties, or groups of images. The ontological language framework proposed here addresses all of these, either through the kind of query terms and predicates used (exact or probabilistic), their explicit or implied scope (either local, compositional, or global), or by means of weighted combination of multiple queries in order to define a partitioning of an image set into groups. In the latter case a composite query could, once defined, be used without further user intervention to categorise or filter new additions to an image collection automatically and one might use a cascade of such queries to impose a hierarchical organisation or derive approximate annotations suited to a particular task.

4.3.2 Query Specific Image Interpretation

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

The use of ontologies also offers the advantage of bridging between high-level concepts and low-level primitives in a way that allows extensions to the language to be defined on the basis of existing constructs without having to alter the representation of image data. Queries can thus span a range of conceptual granularity from concrete image features (regions, colour, shape, texture) and concrete relations (feature distance, spatial proximity, size) to abstract content descriptors (objects, scene descriptions) and abstract relations (similarity, class membership, inferred object and scene composition). The ability to automatically infer the presence of high-level concepts (e.g. a beach scene) on the basis of evidence (colour, region classification, composition) requires techniques such as Bayesian inference, which plays an increasing role in semantic content derivation [281]. By expressing the causal relationships used to integrate multiple sources of evidence and content modalities in a dependency graph, such methods are also of great utility in quickly eliminating improbable configurations and thus narrowing down the search to a rapidly decreasing number of images that are potentially relevant to the query.

4.4 Language Design and Structure

This section introduces the *OQUEL* ontological query language with particular reference to its current implementation as a query description language for the ICON content-based image retrieval system. For reasons of clarity, only a high-level description of the language will be presented here. Section 4.5 will discuss implementation details pertaining to the content extraction and representation schemes used in the system and show how tokens in the language are mapped onto concrete image properties. Section 4.6 will show how query sentences are processed to assess image relevance.

4.4.1 Overview and Design Principles

The primary aim in designing OQUEL has been to provide both ordinary users and professional image archivists with an intuitive and highly versatile means of expressing their retrieval requirements through the use of familiar natural language words and a straightforward syntax.

As mentioned above, many query languages have traditionally followed a path set out by database languages such as SQL, which are characterised by a fairly sparse and restrictive grammatical framework aimed at facilitating concise and well-defined queries. The advantages of such an approach are many, e.g. ease of machine interpretation, availability of query optimisation techniques, scalability, theoretical analysis, etc.. However, their appropriateness and applicability to a domain of such intrinsic ambiguity and uncertainty as image retrieval remains doubtful. OQUEL was therefore designed as a more flexible and natural retrieval tool through the use of a grammar bearing a resemblance to natural language on a restricted domain.

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.

4.4.2 Syntax and Semantics

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

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

While the inherent sophistication of the OQUEL language enables advanced users to specify extremely detailed queries if desired, much of this complexity is hidden by the query parser. The parser was constructed with the aid of the SableCC lexer/parser generator tool from LALR(1) grammar rules and the WordNet [175] lexical database as further described in the next section. The vocabulary of the language is based on an annotated thesaurus of several hundred natural language words, phrases, and abbreviations (e.g. "!" for "not", "," for "and") which are recognised as tokens. Token recognition takes place in a lexical analysis step prior to syntax parsing to reduce the complexity of the grammar. This also makes it possible to provide more advanced word-sense disambiguation and analysis of phrasal structure while keeping the language efficiently LALR(1) parsable.

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

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

The major syntactic categories are:

- *S*: Start symbol of the sentence (text query).
- *R*: Requirement (a query consists of one or more requirements which are evaluated separately, the probabilities of relevance then being combined according to the logical operators).
- *BR*: Binary relation on SBs.
- *SB*: Specification block consisting of at least one CS or PS and 0 or more LS.
- *CS*: Image content specifier.
- *LS*: Location specifier for regions meeting the CS/PS.

- *PS*: Region property specifier (visual properties of regions such as colour, shape, texture, and size).
- *CB*: Binary (fuzzy) logical connective (conjunction, disjunction, and exclusive-OR).

Tokens (terminals) belong to the following sets:

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

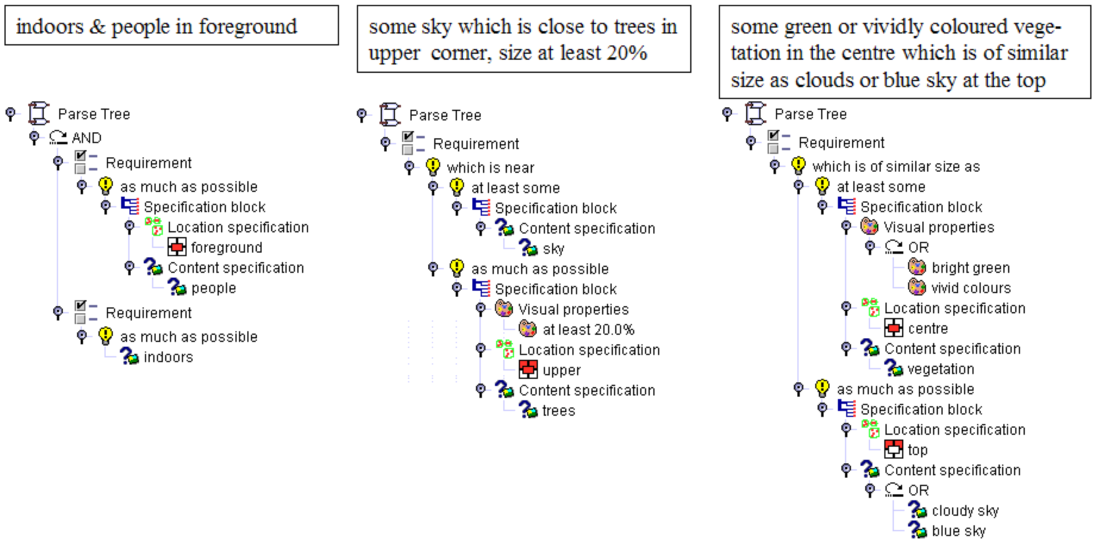
The precise semantics of these constructs are dependent upon the way in which the query language is implemented, the parsing algorithm, and the user query itself, as will be described in the following sections.

4.4.3 Vocabulary

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

In order to provide a rich thesaurus of synonyms and also capture some more complex relations and semantic hierarchies of words and word senses, lexical information from WordNet (see also section 3.1.3) has been utilised. This contains a large vocabulary which has been systematically annotated with word sense information and relationships such as synonyms, antonyms, hyper- and hyponyms, meronyms, etc.. Some of this information was used to define a thesaurus of about 400 words relating to the extracted image features and semantic descriptors mentioned above. However, many of these terms are treated as synonyms by the current implementation (see section 4.5). For example, vision algorithms such as face and skin detectors allow the terms “person” and “people” to be grounded in the image data, but subtypes such as gender (“man”, “woman”) and age (“adult”, “child”) cannot be differentiated at present.

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



indoors & people in foreground

some sky which is close to trees in upper corner, size at least 20%

some green or vividly coloured vegetation in the centre which is of similar size as clouds or blue sky at the top

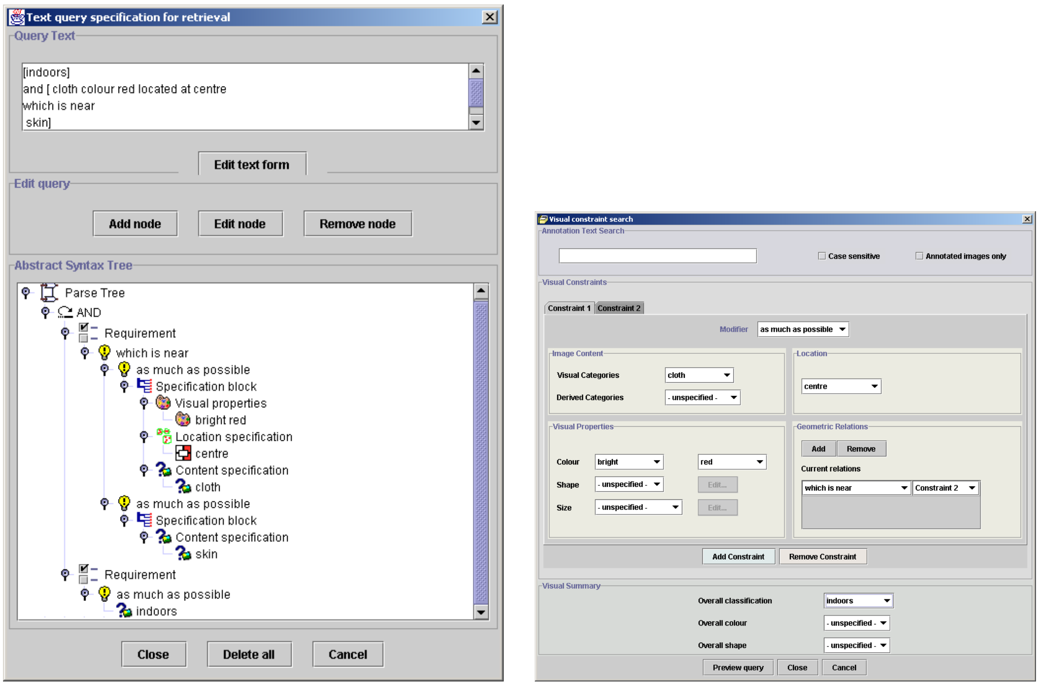
Figure 4.2: Examples of OQUEL query sentences and their syntax trees (as visualised using the ICON application).

4.4.4 Example Sentences

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

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

Figure 4.2 shows some additional query sentences and their resulting abstract syntax trees as visualised using the ICON application (see 4.5). As shown in figure 4.3, such queries may be entered using a simple text dialog or by means of a forms-based graphical user interface.



```
[indoors]
and [ cloth colour red located at centre
which is near
skin]
```

```
Parse Tree
  AND
    Requirement
      which is near
        as much as possible
          Specification block
            Visual properties
              bright red
            Location specification
              centre
            Content specification
              cloth
        as much as possible
          Specification block
            Content specification
              skin
    Requirement
      as much as possible
        indoors
```

□ Case sensitive □ Annotated images only

Modifier: as much as possible

Visual Categories: cloth
Derived Categories: - unspecified -
Location: centre

Colour: bright red
Shape: - unspecified -
Size: - unspecified -

Current relations: which is near Constraint 2

Overall classification: indoors
Overall colour: - unspecified -
Overall shape: - unspecified -

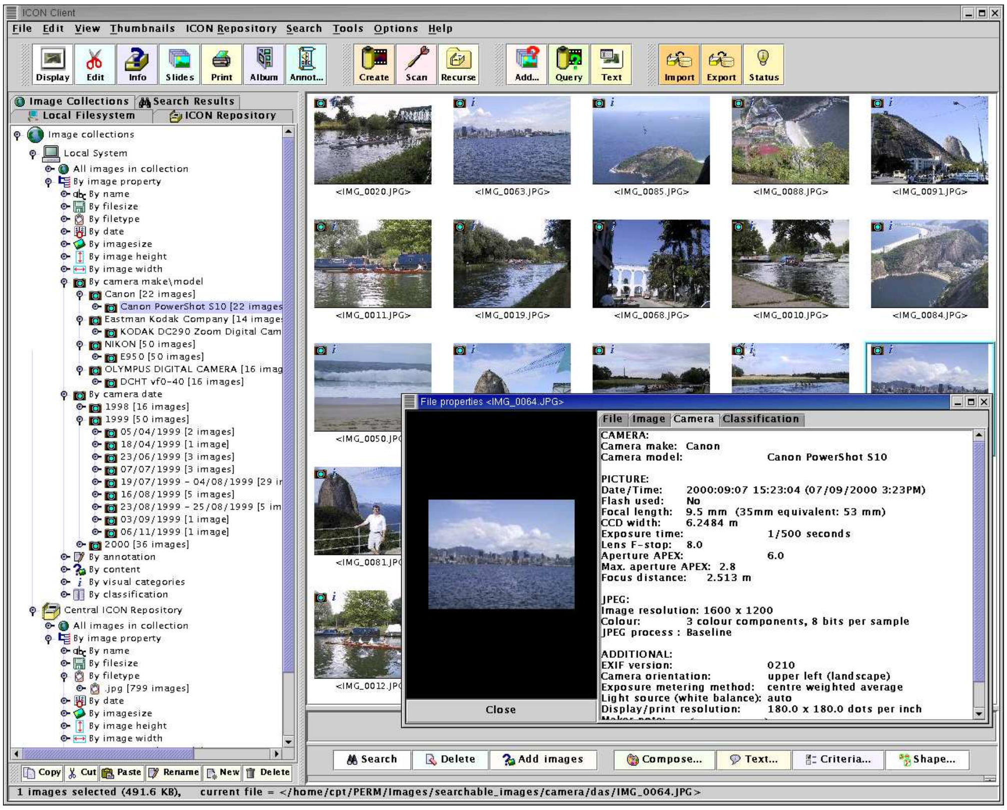
Figure 4.3: Alternative user interfaces representing the same *OQUEL* query.

4.5 Implementation of the OQUEL Language

4.5.1 The ICON System

ICON (Image Content Organisation and Navigation, [269]) combines a cross-platform Java user interface with image processing and content analysis functionality to facilitate automated organisation of and retrieval from large heterogeneous image sets based on both metadata and visual content. As illustrated in figure 4.4, the program allows users to browse images by directory structure and various metadata properties (such as date and camera manufacturer).

In order to adapt to the varying demands of home users, professional photographers, and commercial image collections, the system is designed to be inherently flexible and extensible. A client-server split and object oriented design provide layers of abstraction and encapsulation, thus allowing the system to be customised for a given application domain to meet specific user requirements. ICON currently consists of two parts:



CAMERA:
Camera make: Canon
Camera model: Canon PowerShot S10

PICTURE:
Date/Time: 2000:09:07 15:23:04 (07/09/2000 3:23PM)
Flash used: No
Focal length: 9.5 mm (35mm equivalent: 53 mm)
CCD width: 6.2484 m
Exposure time: 1/500 seconds
Lens F-stop: 8.0
Aperture APEX: 6.0
Max. aperture APEX: 2.8
Focus distance: 2.513 m

JPEG:
Image resolution: 1600 x 1200
Colour: 3 colour components, 8 bits per sample
JPEG process : Baseline

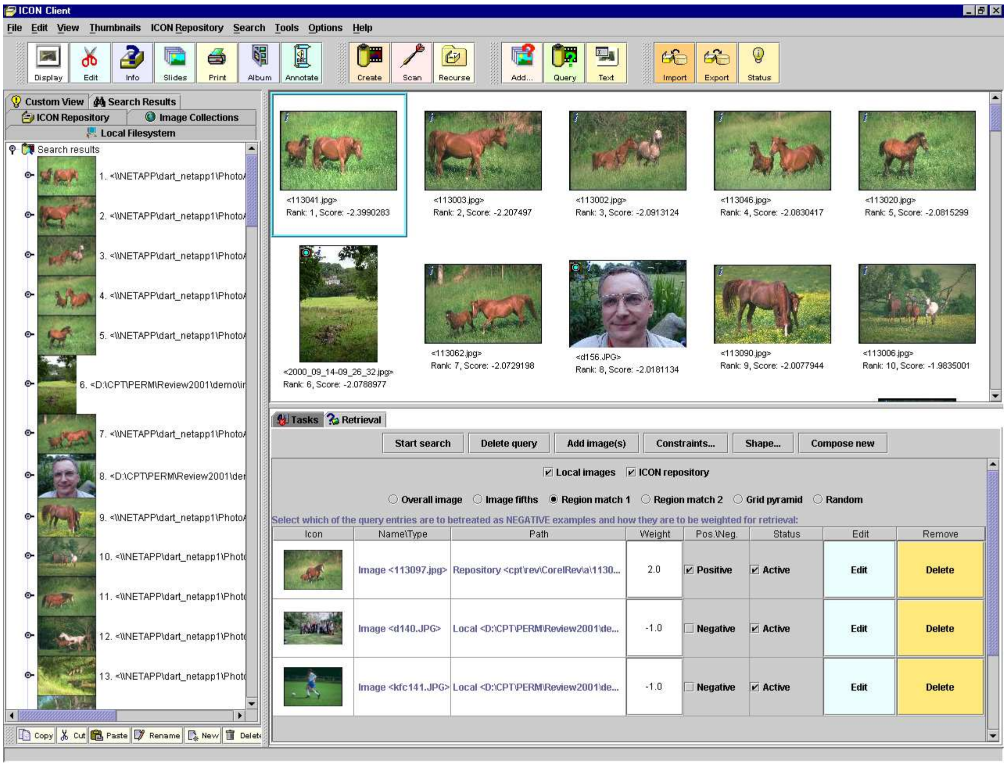
ADDITIONAL:
EXIF version: 0210
Camera orientation: upper left (landscape)
Exposure metering method: centre weighted average
Light source (white balance): auto
Display/print resolution: 180.0 x 180.0 dots per inch

1 images selected (491.6 KB), current file = </home/cpt/PERM/Images/searchable_images/camera/das/IMG_0064.JPG>

Figure 4.4: Screenshot of the ICON image organisation and retrieval interface.

- The ICON client, an application written in Java which can be used either in stand-alone mode or by connecting to the ICON repository. While the current focus is on an exploration of user interface, browsing, and retrieval paradigms for medium-scale collections of personal images, it is envisaged that future client programs will cater for particular application environments such as online image searching and professional photographic archives.

- The ICON repository, a central file archive and set of software tools which apply the segmentation and content classification routines described below to pictures exported from the ICON client. This process enables content-based indexing and searching of images.



1. <\\NETAPP\dart_netapp1\Photo/
2. <\\NETAPP\dart_netapp1\Photo/
3. <\\NETAPP\dart_netapp1\Photo/
4. <\\NETAPP\dart_netapp1\Photo/
5. <\\NETAPP\dart_netapp1\Photo/
6. <D:\CPT\PERM\Review2001\demo\ir
7. <\\NETAPP\dart_netapp1\Photo/
8. <D:\CPT\PERM\Review2001\der
9. <\\NETAPP\dart_netapp1\Photo/
10. <\\NETAPP\dart_netapp1\Photo
11. <\\NETAPP\dart_netapp1\Photo
12. <\\NETAPP\dart_netapp1\Photo
13. <\\NETAPP\dart_netapp1\Photo

<113041.jpg>
Rank: 1, Score: -2.3990283

<113003.jpg>
Rank: 2, Score: -2.207497

<113002.jpg>
Rank: 3, Score: -2.0913124

<113046.jpg>
Rank: 4, Score: -2.0830417

<113020.jpg>
Rank: 5, Score: -2.0815299

<2000_09_14-09_26_32.jpg>
Rank: 6, Score: -2.0788977

<113062.jpg>
Rank: 7, Score: -2.0729198

<d156.JPG>
Rank: 8, Score: -2.0181134

<113090.jpg>
Rank: 9, Score: -2.0077944

<113006.jpg>
Rank: 10, Score: -1.9835001

☑ Local images ☑ ICON repository

○ Overall image ○ Image fifths ● Region match 1 ○ Region match 2 ○ Grid pyramid ○ Random

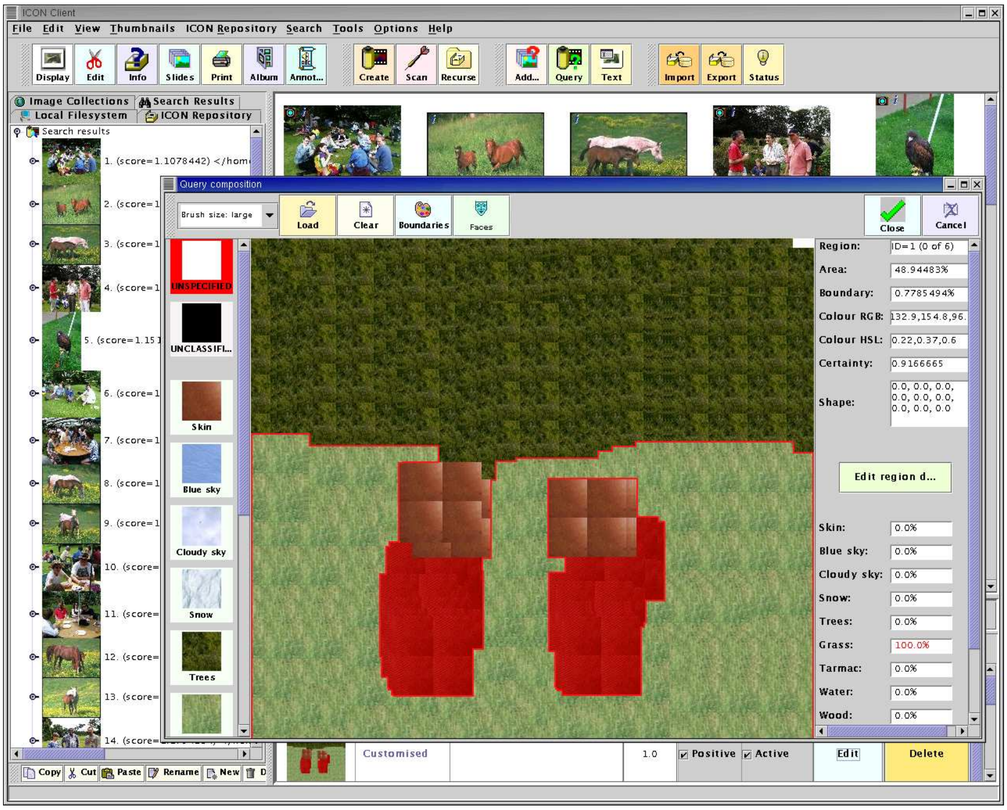
Select which of the query entries are to betreated as NEGATIVE examples and how they are to be weighted for retrieval:

Icon	Name\Type	Path	Weight	Pos.\Neg.	Status	Edit	Remove
	Image <113097.jpg>	Repository <cpt\rev\CorelRev\a\1130...	2.0	☑ Positive	☑ Active	Edit	Delete
	Image <d140.JPG>	Local <D:\CPT\PERM\Review2001\de...	-1.0	☐ Negative	☑ Active	Edit	Delete
	Image <kfc141.JPG>	Local <D:\CPT\PERM\Review2001\de...	-1.0	☐ Negative	☑ Active	Edit	Delete

Figure 4.5: ICON query consisting of a set of positively and negatively weighted example images.

Browsing and organisation

In order to provide a quick and convenient means of viewing and manipulating images via the ICON client, the user is presented with a familiar directory tree view onto the file system. With a single mouse click the client can be made to scan the directory structure to search for images, create thumbnails, and extract metadata such as digital camera settings and annotations. Images can then be exported to the repository to generate visual content descriptors and for archiving purposes. The repository stores metadata and images on a per-user basis but also provides support for collaborative access and for pictures and associated data to be re-imported into the ICON client. Each user may access the repository through an arbitrary number of clients running on any machine or operating system with a current version of the Java runtime environment. Both images stored locally and those in the repository can be browsed and organised according to metadata


ICON Client
File Edit View Thumbnails ICON Repository Search Tools Options Help
Display Edit Info Slides Print Album Annot... Create Scan Recurse Add... Query Text Import Export Status
Image Collections
Search Results
Local Filesystem
ICON Repository
Search results
Query composition
Brush size: large
Load Clear Boundaries Faces
Close Cancel
UNSPECIFIED
UNCLASSIFI...
Skin
Blue sky
Cloudy sky
Snow
Trees
Edit region d...
Copy Cut Paste Rename New D

1. (score=1.1078442) </hom
2. (score=1
3. (score=1
4. (score=1
5. (score=1.15
6. (score=1
7. (score=1
8. (score=1
9. (score=1
10. (score=
11. (score=
12. (score=
13. (score=
14. (score=

Region: ID=1 (0 of 6)
Area: 48.94483%
Boundary: 0.7785494%
Colour RGB: 132.9,154.8,96.
Colour HSL: 0.22,0.37,0.6
Certainty: 0.9166665
Shape: 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0

Skin: 0.0%
Blue sky: 0.0%
Cloudy sky: 0.0%
Snow: 0.0%
Trees: 0.0%
Grass: 100.0%
Tarmac: 0.0%
Water: 0.0%
Wood: 0.0%

Customised 1.0 ☑ Positive ☑ Active Edit Delete

Figure 4.6: Example illustrating the sketch-based ICON query composition interface using a visual thesaurus of semantic categories.

and visual properties rather than just on a per-directory basis. The ICON client provides a range of methods for easy navigation of potentially very large image sets, including clustering and visualisation functionality.

Image retrieval

The image analysis carried out by the ICON repository (see below) segments pictures into regions with associated visual properties and uses neural network classifiers to assign a probabilistic labelling of such image regions with semantic terms corresponding to visual categories such as grass, sky, and water. The ICON client allows image databases to be searched according to metadata (e.g. picture date and digital camera make), annotations, and classified image content. Queries can be formulated in a number of different ways to cater for a range of different retrieval needs and levels of detail in the user's conceptualisation of desired image material. A query may comprise one or several of the following elements:

- A set of weighted sample images which can be either positive or negative (see figure 4.5).
- Desired image content composed my means of a sketch-based query composition tool which uses a visual thesaurus of target image content corresponding to the set of visual categories (see figure 4.6).
- Criteria for various properties including file attributes (e.g. modification date), digital camera settings (e.g. camera model, flash), textual annotations (such as the artist's name of a painting), and constraints on visual appearance features (colour, shape, texture). Figure 4.16 shows an example of such a query.
- A textual or forms-based query expressed in OQUEL (see figure 4.3).

The user may assign different weights to the various elements that comprise a query and can choose from a set of similarity metrics to specify the emphasis that is to be placed on the relative localisation of target content within images and overall compositional aspects.

The backend image processing components extract various types of content descriptors and metadata from images (see [269]). The following sections describe the image analysis processes that are currently used in conjunction with OQUEL queries.

4.5.2 Image Segmentation

Images are segmented[2] into non-overlapping regions and sets of properties for size, colour, shape, and texture are computed for each region [243, 244]. Initially full three colour edge detection is performed using the weighted total change dT

$$dT = dI_i^2 + dI_j^2 + 3.0dC \tag{4.1}$$

[2] The segmentation method used in this work was originally developed by Dr David Sinclair.

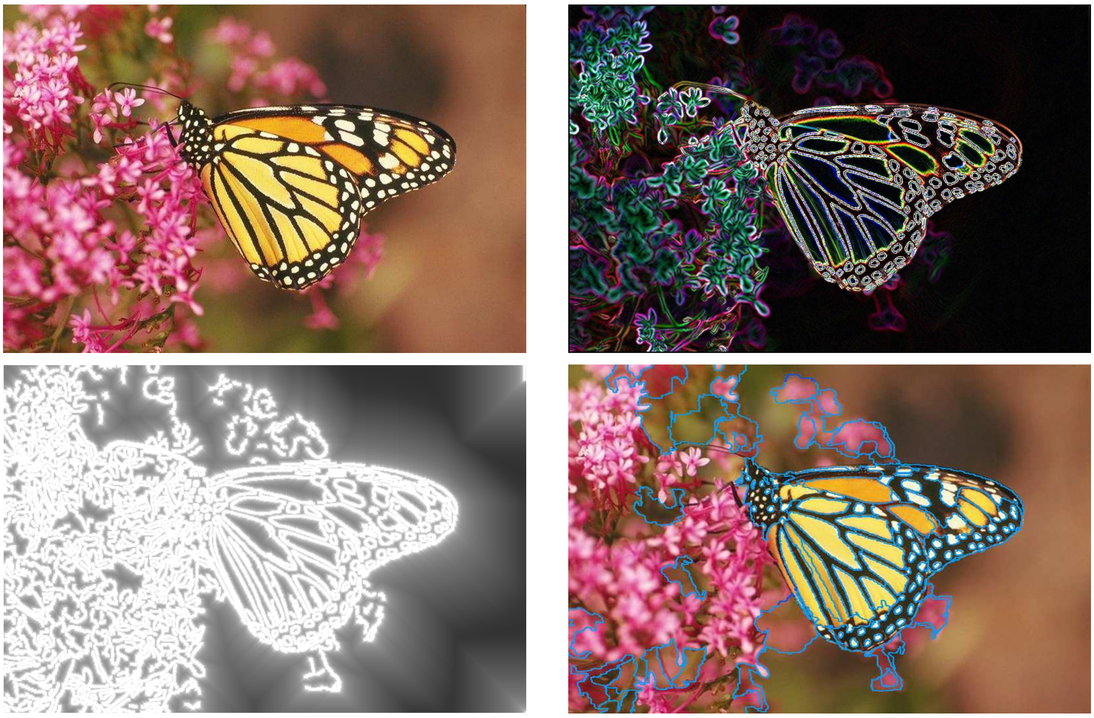


Figure 4.7: From left to right, top to bottom: Example colour image from the Corel image library. Full three colour derivative of the image. Voronoi image computed from the edges found by the three colour edge detector (the darker a pixel, the further it is from an edge). Final region segmentation (boundaries indicated in blue).

where the total change in intensity dI_i along image dimension i is given by the colour derivatives in RGB space

$$dI_i = dR_i + dG_i + dB_i \tag{4.2}$$

and the magnitude of change in colour is represented by

$$\begin{aligned} dC &= [(dB_i - dG_i)^2 + (dR_i - dB_i)^2 + (dG_i - dR_i)^2 \\ &+ (dB_j - dG_j)^2 + (dR_j - dB_j)^2 + (dG_j - dR_j)^2]^{\frac{1}{2}} \end{aligned} \tag{4.3}$$

The choice of 3 as the weighting factor in favour of colour change over brightness change is empirical but has been found to be effective across a very broad range of photographic images and artwork. Local orientation (for use in the non-maximum suppression step of the edge detection process) is defined to be in the direction of the maximum colour gradient. *dT* is then the edge strength fed to the non-max suppression and hysteresis edge-following steps which follow the method due to *Canny* [37].

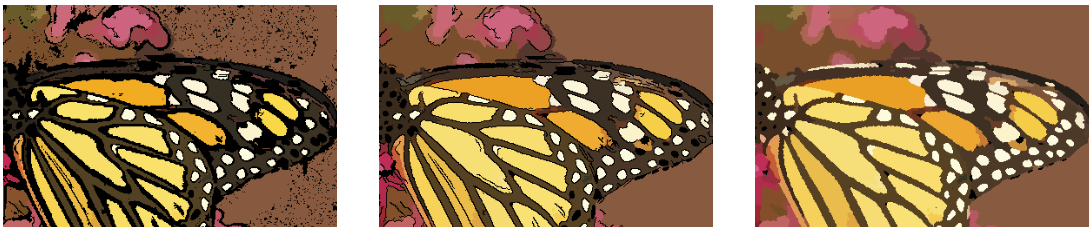


Figure 4.8: Example illustrating the process of region formation using part of the image shown in figure 4.7. ***Left*****: Initial regions grown from the Voronoi image seed points.** ***Middle*****: The thresholds on pixel colour are relaxed and regions grown out to image edges.** ***Right*****: Edges are competitively morphed into adjacent regions and regions of similar colour are merged.**

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

Figure 4.7 shows a sample image from the Corel picture library and the results of segmentation, while figure 4.8 illustrates the process of region formation in greater detail.

A texture model based on discrete ridge features is also used to describe regions in terms of texture feature orientation and density. Ridge pixels are those for which the magnitude of the second derivative operator applied to a grey-scale version of the original image exceeds a threshold. The network of ridges is then broken up into compact 30 pixel feature groups and the orientation of each feature is computed from the second moment about its centre of mass. Feature mean colour is computed from the original image, and connectivity to neighbouring features

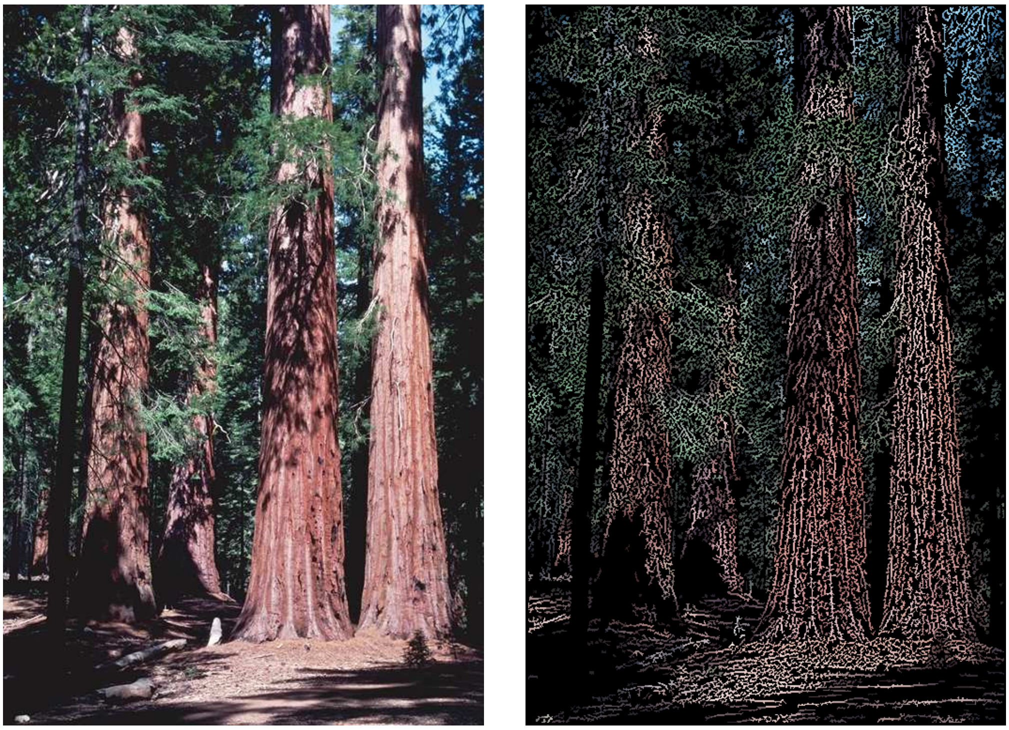


Figure 4.9: Example image and its ridge network broken up into discrete compact texture features. ***Left*****: Corel image;** ***Right*****: Ridge map broken into discrete texture features.**

recorded. The size of features was empirically chosen to be large enough to give reliable estimates of feature colour and shape. Figure 4.9 shows an image and its ridge network broken up into small discrete compact texture features. Net-like structures may be recovered directly from images as a by product of the texture model (see figure 4.10). The topology of connected ridge maps is analysed for holes (simply connected enclosed regions not belonging to the ridge), and if a ridge map has more than five holes it is deemed to be a net.

Features are clustered using Euclidean distance in RGB space and the resulting clusters are then employed to unify regions that share significant portions of the same feature cluster. The list of pair-wise distances between proximate texture features is computed and ordered by increasing distance. Each pair-wise relation is used in turn to start a texture cluster centre (here referred to as a clique), to add a feature to a clique, or to amalgamate two existing cliques. Internal texture structure of a textured region can then be described by an orientation

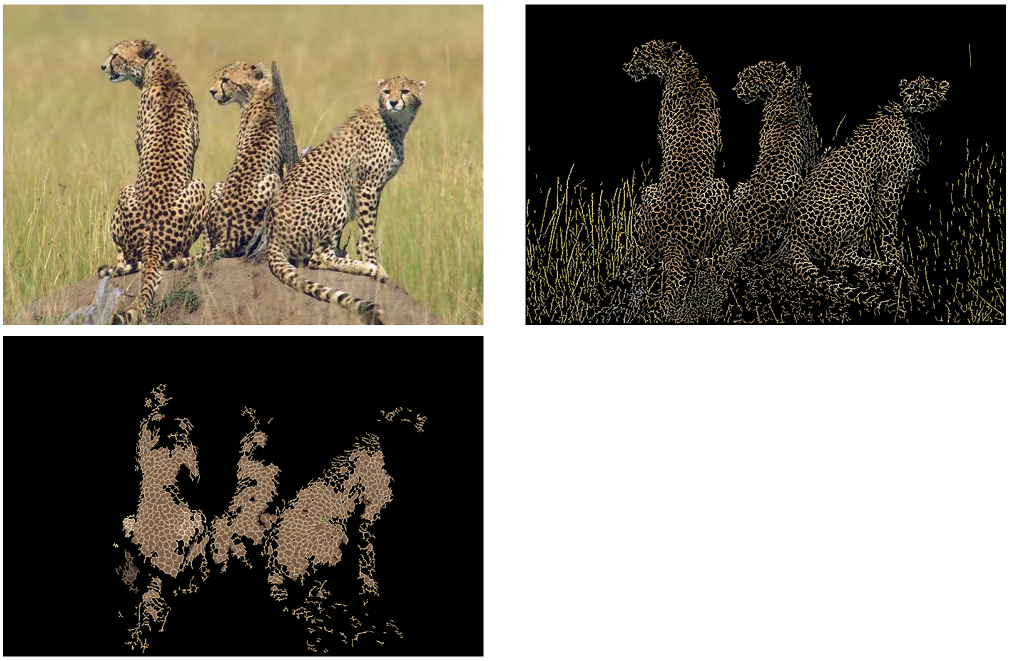


Figure 4.10: Example image, ridge network, and connected net structures of texture features. *Top left*: **Example Corel image;** *Top right*: **Ridge map broken into discrete texture features;** *Bottom*: **Connected net structures extracted from the ridge map.**

histogram of the texture features and a pair-wise geometric histogram [121] of relative orientation versus mutual separation.

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 [244]. A histogram representation of colour covariance and shape features is computed for regions above a certain size threshold.

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

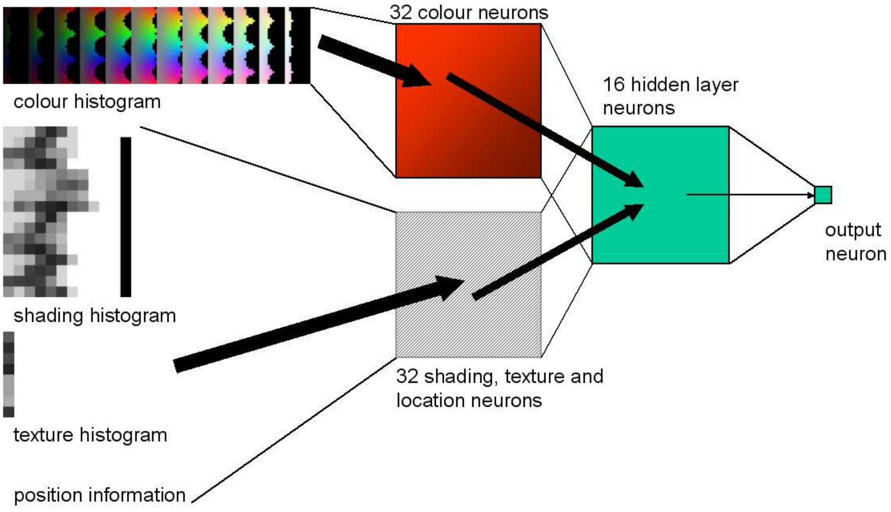


Figure 4.11: Example architecture of the Multi-layer Perceptron neural networks used for image region classification.

4.5.3 Region Classification

Region descriptors computed from the segmentation algorithm are fed into artificial neural network classifiers which have been trained to label regions with class membership probabilities for a set of 12 semantically meaningful visual categories of "stuff" ("Brick", "Blue sky", "Cloth", "Cloudy sky", "Grass", "Internal walls", "Skin", "Snow", "Tarmac", "Trees", "Water", and "Wood").

The classifiers are MLP (Multi-layer Perceptron) and RBF (Radial Basis Function) networks whose topology was optimised to yield best generalisation performance for each particular visual category using separate training and validation sets from a large (over 40000 exemplars) corpus of manually labelled image regions. The MLP networks typically consist of two or three hidden layers with

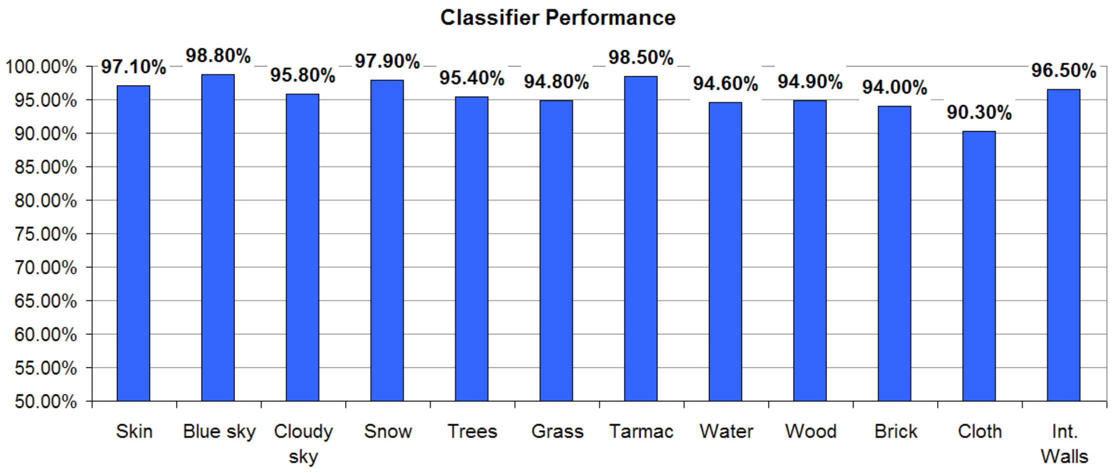


Figure 4.12: Correct classification (true positive) rates achieved by the neural network region classifiers.

progressively smaller numbers of neurons in each layer. This effectively partitions the mapping from input space to output space into several stages of decreasing complexity. Secondly, feature extraction is made more explicit by implementing the notion of *receptive fields*. Instead of connecting all neurons in one layer to all neurons in the next layer (as is usually the case in feed-forward networks), subsets of the hidden layer are connected only to semantically related groupings of region descriptors. This introduces a neighbourhood relation between neurons of the same layer, especially if the receptive field areas are made to overlap according to the proximity of corresponding neurons. After training, the continuous output of the neural networks (which represents a class membership probability) is thresholded using a value of 0.6 to arrive at a binary classification. Figure 4.11 shows an example of the MLP network structure.

Automatic labelling of segmented image regions with semantic visual categories [269] such as grass or water which mirror aspects of human perception allows the implementation of intuitive and versatile query composition methods while greatly reducing the search space. The current set of categories was chosen to facilitate robust classification of general photographic images. These categories are by no means exhaustive but represent a first step towards identifying fairly low-level semantic properties of image regions which can be used to ground higher level

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

concepts and content prescriptions. Various psychophysical studies [53, 178] have shown that semantic descriptors such as these serve as useful cues for determining image content by humans and CBIR systems. An attempt was made to include categories that allow one to distinguish between indoor and outdoor scenes.

In addition to the training and validation sets, classifier performance was tested using a separate test set of 1436 randomly selected image regions. As shown in figure 4.12, the classifiers achieve classification success rates of between 90.3% and 98.8%. Although these figures are impressive and indicate a high overall accuracy and low false positive rate, it is also important in practice to ensure separation between different classes. Evaluation results from the test set were used to obtain the classifier confusion matrix shown in table 4.1. 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.

Clearly the matrix is not symmetric, for example the probability of a region of skin being mis-classified as wood is estimated to be 12%, whereas the probability of a wood region being mis-classified as skin over the testing set is only 2%. Clearly there are instances of different classes that are intrinsically difficult to discriminate due to their visual similarity, for example regions of trees at particular scales may appear very similar to grass and some notion of scene context may be required to achieve successful disambiguation. Likewise, the large intra-class appearance variation of categories such as cloth and water under different lighting and viewing conditions means that they will have some overlap with other classes. Unresolvable ambiguities may therefore arise in the classification of particular image regions.

As will be shown below, one of the strengths of the OQUEL retrieval framework lies in the fact that it is resilient with respect to such errors and ambiguities in the image content classification. The structure of the ontology and the probabilistic nature of query evaluation and matching make it possible to exploit the syntactic and semantic redundancies in the image set.

4.5.4 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 [192, 178], colour classifiers were constructed for each of twelve "basic" colours ("black", "blue", "brown", "cyan", "green", "grey", "magenta", "orange", "pink", "red", "white", "yellow"). Each region is associated with the colour labels which best describe it (see section 4.5.7).

4.5.5 Face Detection

The face detector[3] 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 [230] 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 pair-wise geometric histogram binning

[3] This face detection method was originally developed by Dr David Sinclair.

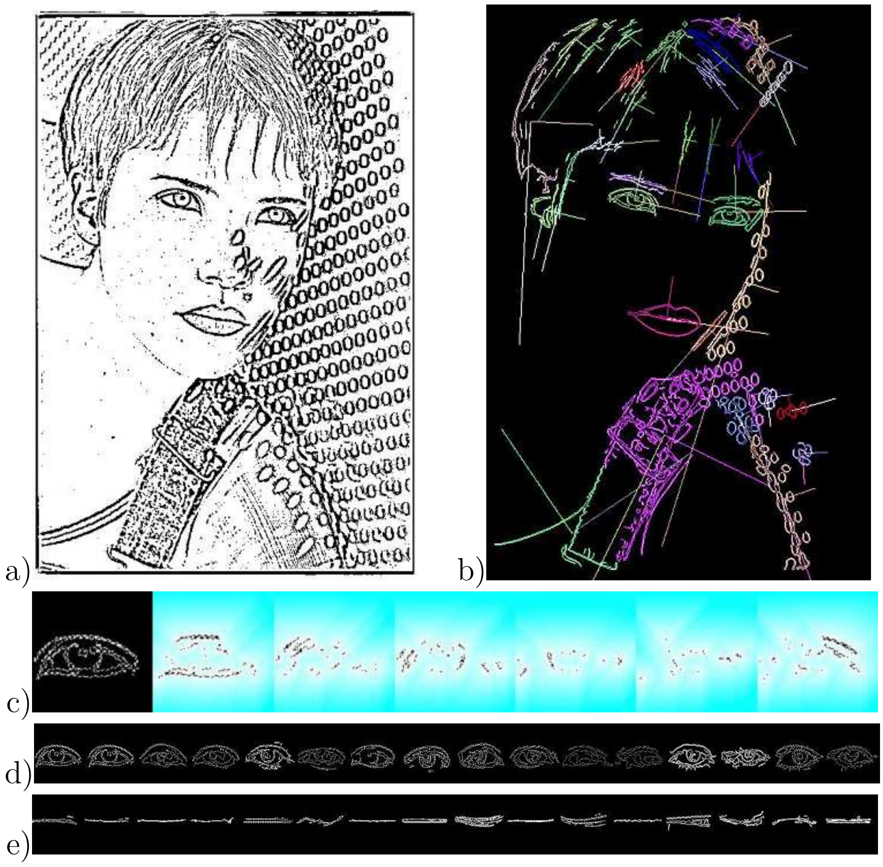


Figure 4.13: Eye detection method employed by the ICON system: a) Binarised image. b) Hausdorff clustered regions after filtering. c) Normalised feature cluster of the left eye (left) and distance transforms for 6 feature orientations (blue areas are further from feature points). d) Examples of left eyes correctly classified using nearest neighbours. e) Examples of nearest neighbour clusters for non-eyes.

[77]. A nearest-neighbour shape classifier was trained to recognise eyes. See figure 4.13 for an illustration of the approach.

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

4.5.6 Image Content Representation

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

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

- *Region graph*: Graph of the relative spatial relationships of the regions (distance, adjacency, joint boundary, and containment). Distance is defined in terms of the Euclidean distance between centres of gravity, adjacency is a binary property denoting that regions share a common boundary segment, and the joint boundary property gives the relative proportion of region boundary shared by adjacent regions. A region R1 is said to be contained within region R2 if R1 shares 100% of its boundary with R2. Together with the simple parameterisation of region shape computed by the segmentation method, this provides an efficient (if non-exact) representation of the geometric relationships between image regions.

- *Grid pyramid*: The proportion of image content which has been positively classified with each particular label (visual category, colour, and presence of faces) at different levels of an image pyramid (whole image, image fifths, 8x8 chess grid, see figure 4.14). 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. Grid regions are generally of the same area and rectangular shape, except in the case of the image fifths where the central rectangular fifth occupies 25% of image area and is often given a higher weighting for scene characterisation to reflect the fact that this region is likely to constitute the most salient part of the image.

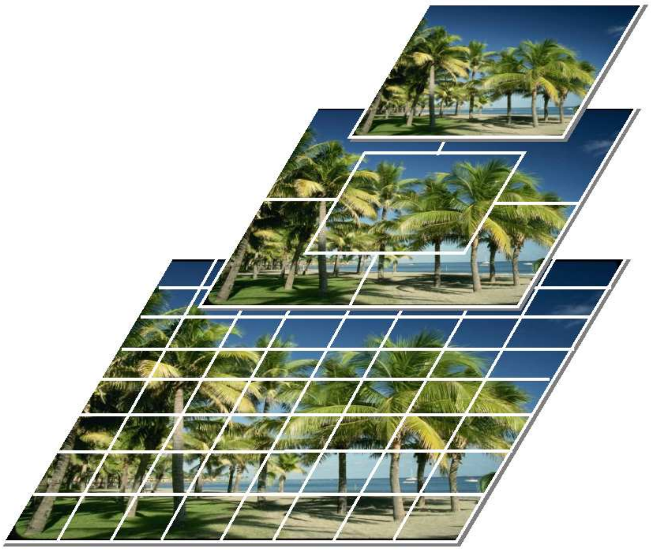


Figure 4.14: Three level grid pyramid which subdivides an image into different numbers of fixed polygonal regions (1, 5, 64) at each level.

Through the relationship graph representation, matching of clusters of regions is made invariant with respect to displacement and rotation using standard matching algorithms [205]. The grid pyramid and region mask representations allow an efficient comparison of absolute position and size.

This may be regarded as an intermediate level representation which does not preclude additional stages of visual inference and composite object recognition in light of query specific saliency measures and the integration of contextual information. Such intermediate level semantic descriptors for image content have been used by several CBIR systems in recent years ([159], [36], [86], [132]).

4.5.7 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 4.6.

- *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 that 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").

4.5.8 System Integration

A general query methodology for content-based image and multimedia retrieval must take into account the differences in potential application domains and system environments. Great care was therefore taken in the design of the OQUEL language to make it possible to integrate it with existing database infrastructure and content analysis facilities. This *portability* was achieved by a component-based software development approach which allows individual matching modules to be re-implemented to cater for alternative content representation schemes. This facility also makes it possible to evaluate a particular query differently depending on the current retrieval context.

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

As will be discussed in the next section, the query text parser was designed to hide some of the grammatical complexity and provide a natural tool for query composition. There is also a forms-based interface which offers the look and feel of graphical database interfaces and explicitly exposes available language features while being slightly restricted in the type of queries it can handle (see figure 4.3). Lastly there is a graphical tool which allows users to inspect or modify a simplified abstract syntax tree (AST) representation of a query.

4.6 Retrieval Process

This section discusses the OQUEL retrieval process as implemented in the ICON system. In the first stage, the syntax tree derived from the query is parsed top-down and the leaf nodes are evaluated in light of their predecessors and siblings. Information then propagates back up the tree until 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.

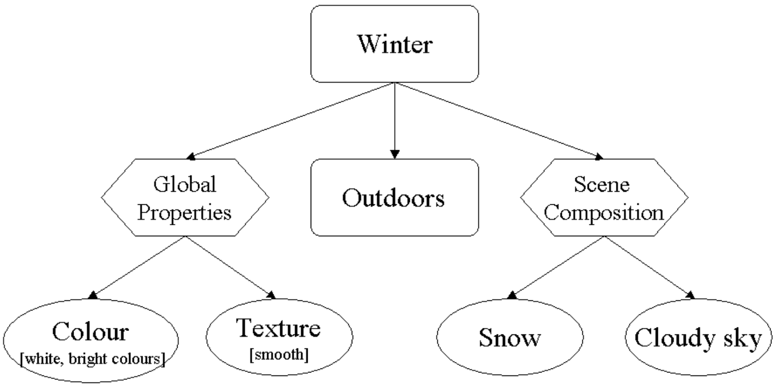


Figure 4.15: Simplified Bayesian network for the scene descriptor "winter".

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.

Figure 4.15 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.

4.6.1 Query-time Object and Scene Recognition for Retrieval

Going back to the lessons learned from text retrieval stated in section 4.2.2, for most content retrieval tasks it is perfectly adequate to approach the problem of retrieving images containing particular objects or characterisable by particular scene descriptors in an *indicative* fashion rather than a full *analytic* one. As long as the structure of the inference methods adequately accounts for the non-accidental properties that characterise an object or scene, relevance can be assessed by a combination of individually weak sources of evidence. These can be ranked in a hierarchy and further divided into those that are *necessary* for the object to be deemed present and those that are merely *contingent*. Such a ranking makes it possible to quickly eliminate highly improbable images and narrow down the search window.

Relevant images are those where 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 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 images where the confidence of having found evidence for the presence of people is high relative to the other images, subject to the inevitable thresholding and identification of necessary features.

4.6.2 Query Parsing and Representation

OQUEL queries are parsed to yield a canonical abstract syntax tree (AST) representation of their syntactic structure. Figures 4.2, 4.17, 4.18, 4.19, and 4.20 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") that represent the relationships that are to be applied when evaluating the query.

4.6.3 Query Evaluation and Retrieval

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

The overall approach therefore relies on passing messages (image structures labelled with probabilities of relevance), assigning weights to these messages according to higher level structural nodes (modifiers and relations), and integrating these at the topmost levels (specification blocks) in order to compute a belief state for the relevance of the evidence extracted from the given image for the given query. There are many approaches to using probabilities to quantify and combine uncertainties and beliefs in this way [196]. The approach adopted here is related to that of [152] in that it applies notions of weighting akin to the Dempster-Shafer theory of evidence to construct an information retrieval model that captures structure, significance, uncertainty, and partiality in the evaluation process.

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

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 [92]. 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.

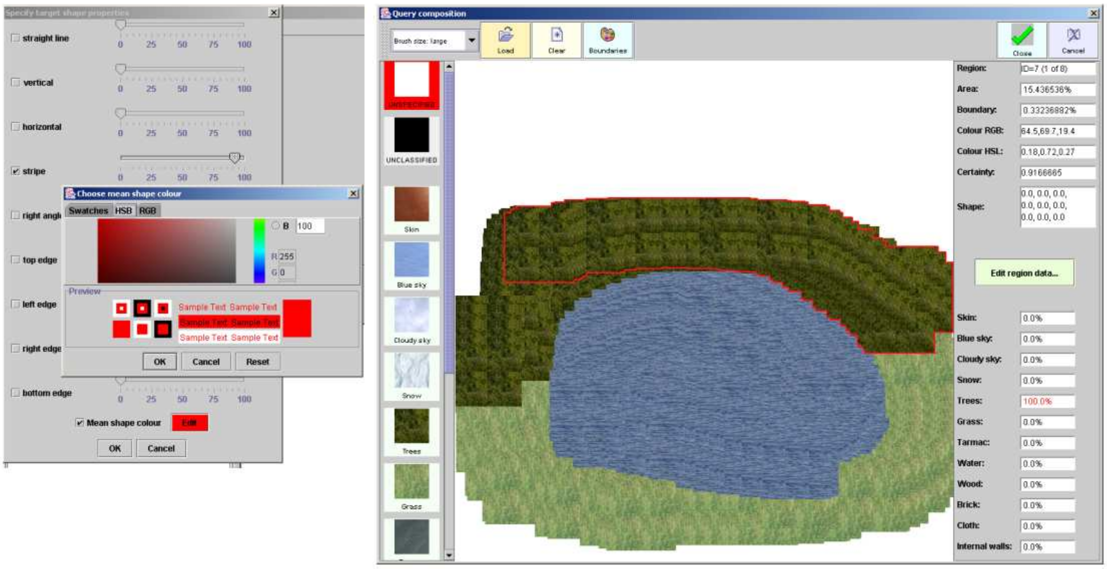
4.7 Evaluation

4.7.1 Qualitative and Quantitative Evaluation

Progress in CBIR research remains hampered by a lack of standards for comparative performance evaluation [181, 180, 284, 60]. This is an intrinsic problem due to the extreme ambiguity of visual information with respect to human vs. computer interpretations of content and the strong dependence of relevance assessment upon the particular feature sets and query methods implemented by a given system. There are no publicly available image sets with associated ground truth data at the different levels of granularity required to do justice to different retrieval approaches, nor are there any standard sets of queries and manually ranked results that could easily be translated to the different formats and conventions adopted by different CBIR systems.

Furthermore, there are no usable automated techniques for assessing important yet elusive usability criteria relating to the query interface as discussed in 4.2.1. Real-world users (rarely addressed in the CBIR research literature) would be primarily interested in the ease with which they could formulate effective queries in a particular system to solve their search requirements with minimal effort for their chosen data set. Even if large scale standardised test sets and sample queries were available to the CBIR community, results derived from them might not be of much use in predicting performance on real-world retrieval tasks.

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



□ straight line

□ vertical

□ horizontal

☑ stripe

□ right angle

□ top edge

□ left edge

□ right edge

□ bottom edge

☑ Mean shape colour

○ B 100

R 255

G 0

Region: ID=7 (1 of 8)

Area: 15.436536%

Boundary: 0.33236882%

Colour RGB: 64.5,69.7,19.4

Colour HSL: 0.18,0.72,0.27

Certainty: 0.9166665

Shape: 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0

Skin: 0.0%

Blue sky: 0.0%

Cloudy sky: 0.0%

Snow: 0.0%

Trees: 100.0%

Grass: 0.0%

Tarmac: 0.0%

Water: 0.0%

Wood: 0.0%

Brick: 0.0%

Cloth: 0.0%

Internal walls: 0.0%

Figure 4.16: Examples of alternate ICON query interfaces using region properties (left) and sketch of classified target content (right).

4.7.2 Experimental Method

While most evaluation of CBIR systems is performed on commercial image collections such as the Corel image sets, their usefulness is limited by the fact that they consist of very high quality photographic images and that the associated ground truth (category labels such as "China", "Mountains", "Food") are frequently too high-level and sparse to be of use in performance analysis [180]. Therefore a set of images consisting of around 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. In the case of the four test queries A, B, C, and D introduced in section 4.7.3 below, the number of relevant images was 77, 158, 53, and 67 respectively.

In order to quantify the performance of the current implementation of the OQUEL language in light of the inherent difficulties of CBIR evaluation, the evaluation focuses on contrasting its utility as a retrieval tool compared with the other query modalities present in the ICON system. As mentioned in section 4.5.1, these are:

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

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

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

As discussed above, the user may assign different weights to the various elements that comprise a query and can choose from a set of similarity metrics to specify the emphasis that is to be placed on the absolute position of target content within images and overall compositional aspects. All of the query components have access to the same pre-computed image representation as described in 4.5.

4.7.3 Test Queries

Four test queries were chosen, which have the following expressions in the OQUEL language:

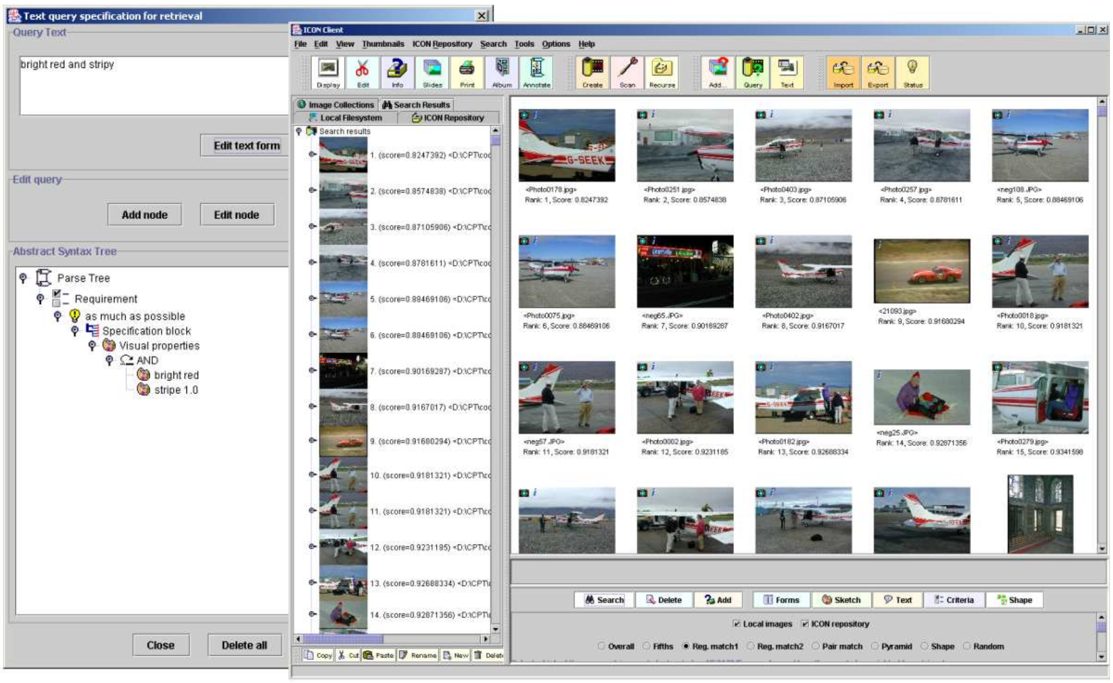
- *Query A* “bright red and stripy”

- *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”

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

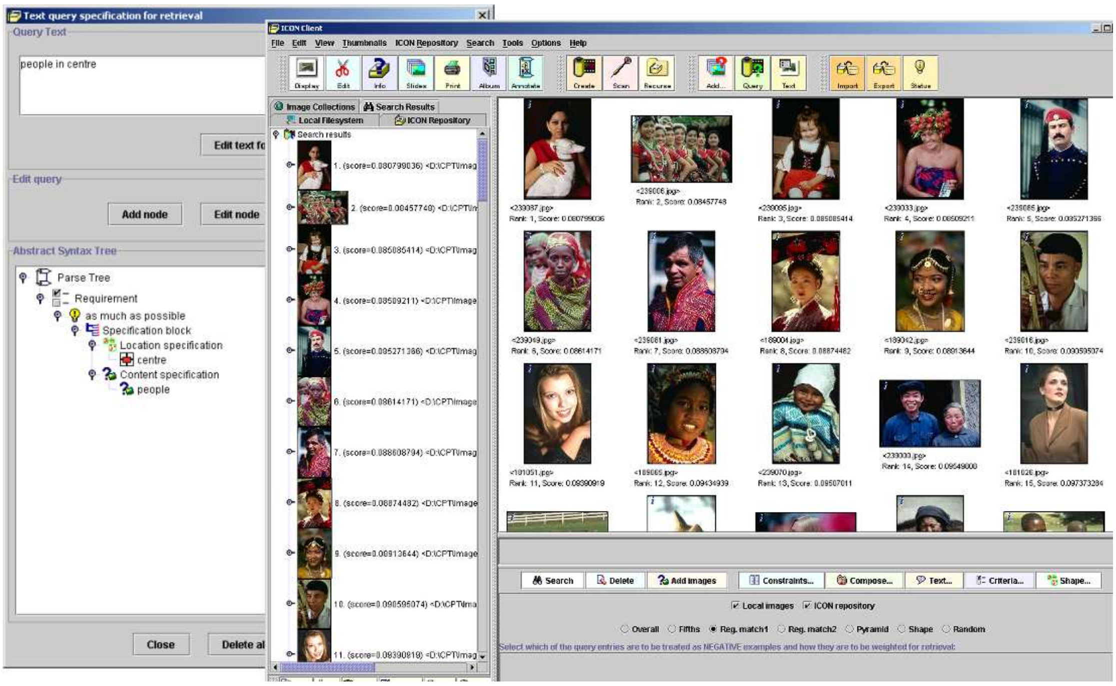

Text query specification for retrieval
Query Text
Edit text form
Edit query
Add node
Edit node
Abstract Syntax Tree
Close
Delete all
ICON Client
File Edit View Thumbnails ICON Repository Search Tools Options Help
Display Edit Info Slides Print Album Annotate Create Scan Recurse Add.. Query Text Import Export Status
Image Collections
Search Results
Local Filesystem
ICON Repository
Copy Cut Paste Rename New Delete
Search Delete Add Forms Sketch Text Criteria Shape
☑ Local images ☑ ICON repository
○ Overall ○ Fifths ● Reg. match1 ○ Reg. match2 ○ Pair match ○ Pyramid ○ Shape ○ Random

bright red and stripy

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

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

Figure 4.17: Search results for OQUEL query A "bright red and stripy".


Text query specification for retrieval
Query Text
Edit text fo
Edit query
Add node
Edit node
Abstract Syntax Tree
Close
Delete al
ICON Client
File Edit View Thumbnails ICON Repository Search Tools Options Help
Display Edit Info Slides Print Album Annotate Create Scan Recurse Add.. Query Text Import Export Status
Image Collections
Search Results
Local Filesystem
ICON Repository
Search Delete Add images Constraints... Compose... Text... Criteria... Shape...
☑ Local images ☑ ICON repository
○ Overall ○ Fifths ● Reg. match1 ○ Reg. match2 ○ Pyramid ○ Shape ○ Random

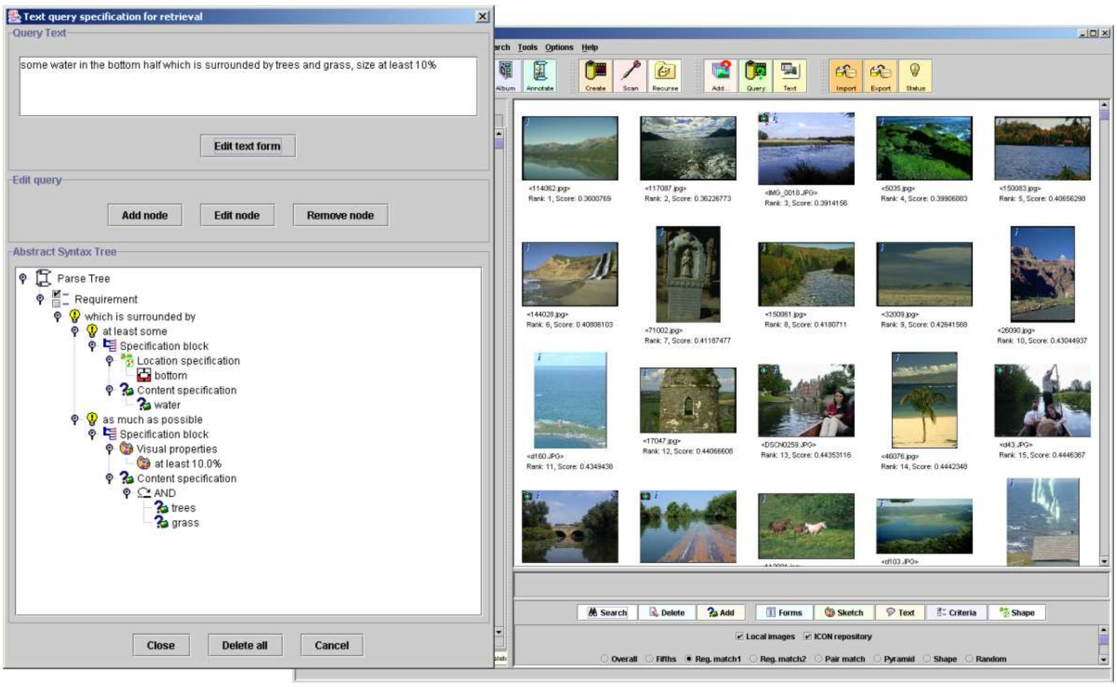
people in centre

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

Search results
1. (score=0.080799036) <D:\CPT\imag
2. (score=0.08457748) <D:\CPT\im
3. (score=0.085085414) <D:\CPT\imag
4. (score=0.08509211) <D:\CPT\image
5. (score=0.085271386) <D:\CPT\imag
6. (score=0.08614171) <D:\CPT\image
7. (score=0.088608794) <D:\CPT\imag
8. (score=0.08874482) <D:\CPT\image
9. (score=0.08913644) <D:\CPT\image
10. (score=0.090595074) <D:\CPT\ima
11. (score=0.09390819) <D:\CPT\imag

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

Figure 4.18: Search results for OQUEL query B "people in centre".



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

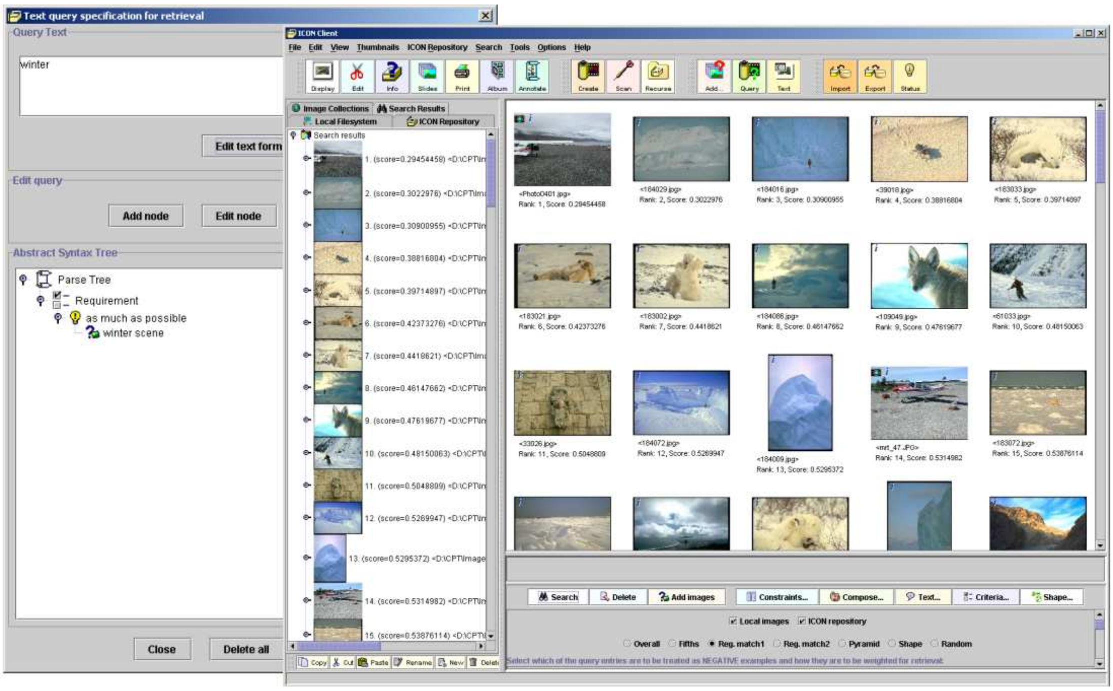
Parse Tree
Requirement
which is surrounded by
at least some
Specification block
Location specification
bottom
Content specification
water
as much as possible
Specification block
Visual properties
at least 10.0%
Content specification
AND
trees
grass

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

☑ Local images ☑ ICON repository

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

Figure 4.19: Search results for OQUEL query C "some water in the bottom half which is surrounded by trees and grass, size at least 10%".



winter

Parse Tree
Requirement
as much as possible
winter scene

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

<Photo0401.jpg> Rank: 1, Score: 0.29454458
<184029.jpg> Rank: 2, Score: 0.3022976
<184016.jpg> Rank: 3, Score: 0.30900955
<39018.jpg> Rank: 4, Score: 0.38816804
<183033.jpg> Rank: 5, Score: 0.39714897
<183021.jpg> Rank: 6, Score: 0.42373276
<183002.jpg> Rank: 7, Score: 0.4418621
<184086.jpg> Rank: 8, Score: 0.46147662
<109049.jpg> Rank: 9, Score: 0.47619677
<61033.jpg> Rank: 10, Score: 0.48150063
<33026.jpg> Rank: 11, Score: 0.5048809
<184072.jpg> Rank: 12, Score: 0.5269947
<184009.jpg> Rank: 13, Score: 0.5295372
<mt_47.JPG> Rank: 14, Score: 0.5314982
<183072.jpg> Rank: 15, Score: 0.53876114

☑ Local images ☑ ICON repository

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

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

Figure 4.20: Search results for OQUEL query D "winter".

for conceptually different retrieval needs expressed at different levels of description. For each OQUEL query a further two queries embodying the same retrieval need were expressed using the other search facilities of the ICON system:

- Combined query: a query that may combine a sketch with feature constraints as appropriate to yield best performance in reasonable time.

- Query-by-example: the single image maximising the normalised average rank metric was chosen as the query. This type of query is commonly used to assess baseline performance.

Figures 4.17, 4.18, 4.19, and 4.20 show the four OQUEL queries and their search results over the collection. Figure 4.16 depicts examples of alternate queries consisting of a combination of low-level attributes and user drawn sketch.

4.7.4 Results

To quantify performance, graphs of precision versus recall and number of images retrieved versus relevant images were computed using the ground truth images for each test query as shown in figure 4.21. For each OQUEL query results are also shown 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 as the corresponding OQUEL query. It can be seen that OQUEL queries yield better results, especially for the top ranked images. In the case of query A, results are essentially the same as those for a query consisting of feature predicates for the region properties "stripy" and "red". In general OQUEL queries are more robust to errors in the segmentation and region classification due to their ontological structure. Query-by-example in particular is usually insufficient to express more advanced concepts relating to spatial composition, feature invariances, or object level constraints.

As recommended in [181], the *normalised average rank* is also computed, which is a useful stable measure of relative performance in CBIR:

$$Rank^{\sim} = \frac{1}{NN_{rel}} \left[\sum_{i=1}^{N_{rel}} R_i - \frac{N_{rel}\left(N_{rel}+1\right)}{2} \right] \tag{4.4}$$

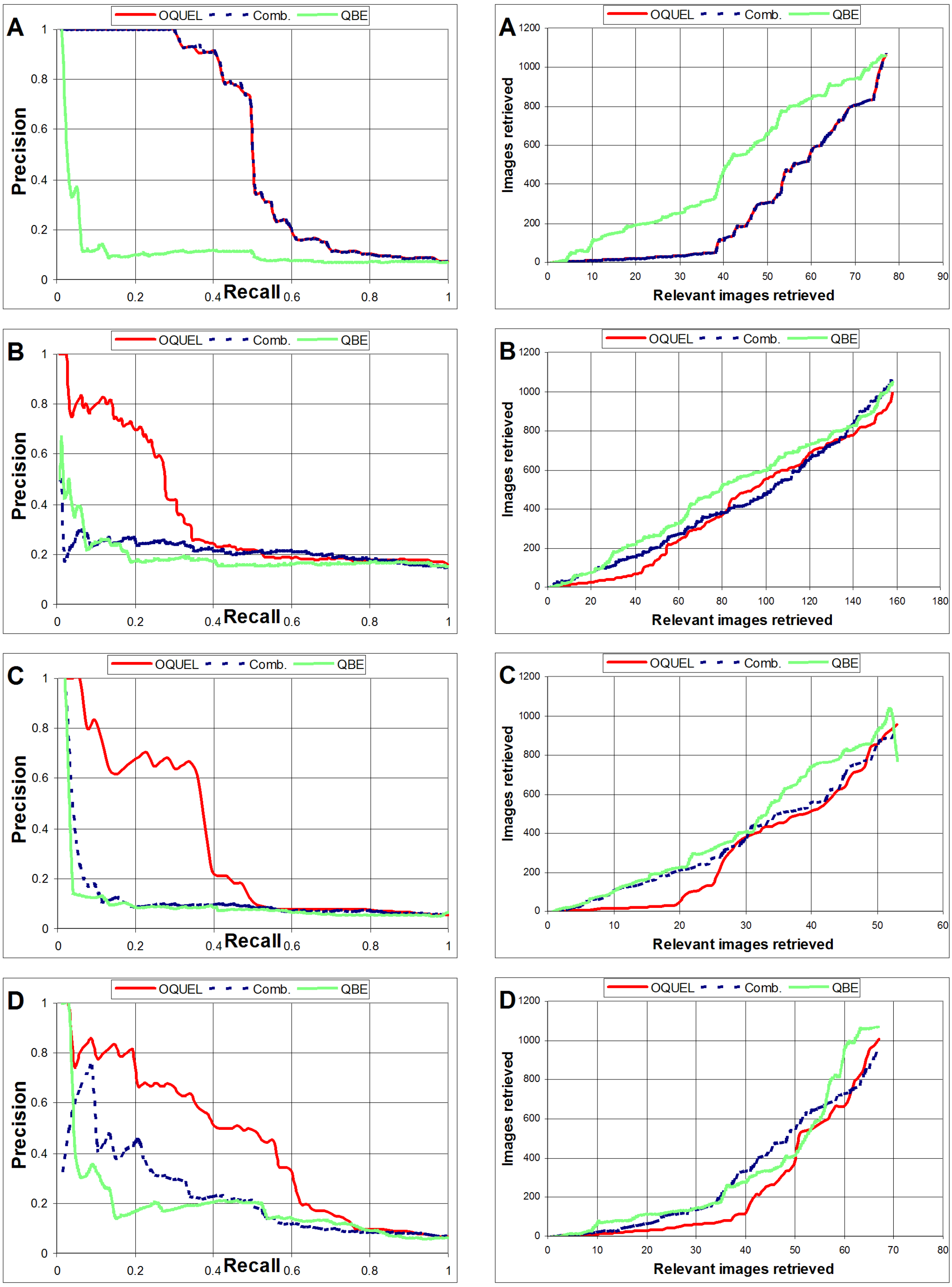


Figure 4.21: Plots of *left*: precision versus recall and *right*: total number of images retrieved versus number of relevant images retrieved for the 4 retrieval experiments (A, B, C, D) as computed for each of the three query modalities.

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.

Query	$Rank^{\sim}$
A - OQUEL	0.2176
A - Comb.	0.2175
A - QBE	0.3983
B - OQUEL	0.2915
B - Comb.	0.3072
B - QBE	0.3684
C - OQUEL	0.2628
C - Comb.	0.3149
C - QBE	0.3521
D - OQUEL	0.1935
D - Comb.	0.2573
D - QBE	0.2577

Table 4.2: Results of the query experiments indicating the normalised average rank measure for each of 4 query experiments (A, B, C, D) and for three methods of query composition (OQUEL, "combined", and "query-by-example").

Comparisons with other query composition and retrieval paradigms implemented in ICON (sketch, sample images, property thresholds) therefore show that the OQUEL query language constitutes a more efficient and flexible retrieval tool (see table 4.2). 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. Text queries have also generally proven to be more efficient to evaluate. This is because 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. Although the current system does not use an inverted file as its index, query evaluation took no more than 100ms for the test queries.

4.7.5 Scalability Experiment

One important aspect of any retrieval technology is whether it scales well with respect to the ever growing size of data sets. This is particularly important in image retrieval, where an increase in magnitude from e.g. 1000 to 10000 images is likely to mark the transition from relatively small, well known personal image collections to much larger data sets with which no individual person is familiar, thus providing the *raison d'être* for content-based image retrieval solutions.

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 (although content-based methods for detecting and rectifying photographic image orientation exist, e.g. [279, 287]).The image analysis methods described in section 4.5 were applied to index the entire collection without pre-processing or altering its contents in any way.

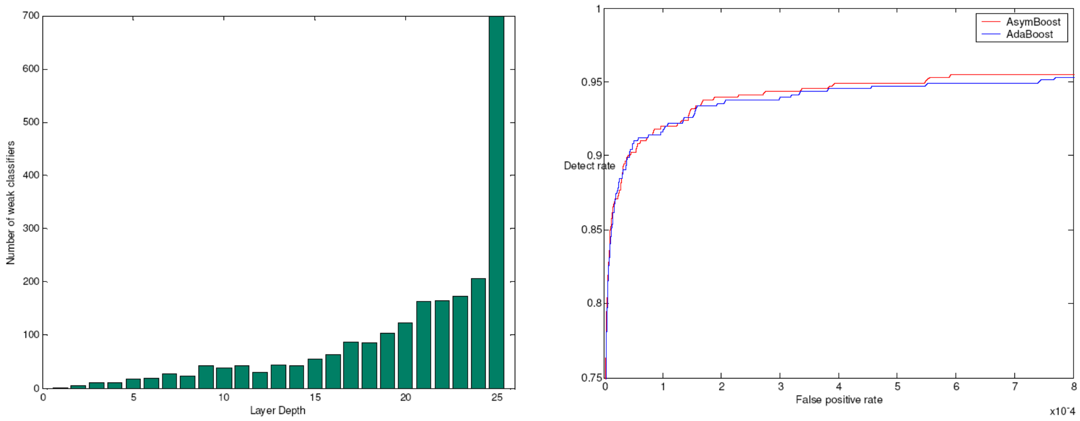


Figure 4.22: Structure and performance results for a face detection method consisting of a cascade of simple feature classifiers. ***Left*****: Number of component classifier features at each level of the cascade.** ***Right*****: Receiver Operating Characteristic (ROC) curves for face detector cascades trained using standard (blue) and asymmetric (red) AdaBoost.**

Although most of the image analysis takes only a few seconds for even very high-resolution images, the face detection process explained in section 4.5.5 proved somewhat prohibitive. Consequently a more efficient algorithm based on the popular face detection method by Viola and Jones [283] was implemented. The method makes use of the AdaBoost learning algorithm to train a cascade of face classifiers consisting of simple bar features. The response of these features can be computed very rapidly and only those image subregions that are not rejected by earlier layers in the cascade have to be considered by the more complex classifiers at higher layers. Consequently the goal of the learning algorithm is to create classifiers which exhibit extremely low false reject rates at moderate false positive rates to ensure both high efficiency and accuracy, since the detector is typically applied to very large numbers of rectangular candidate regions at different positions and scales in a given image.

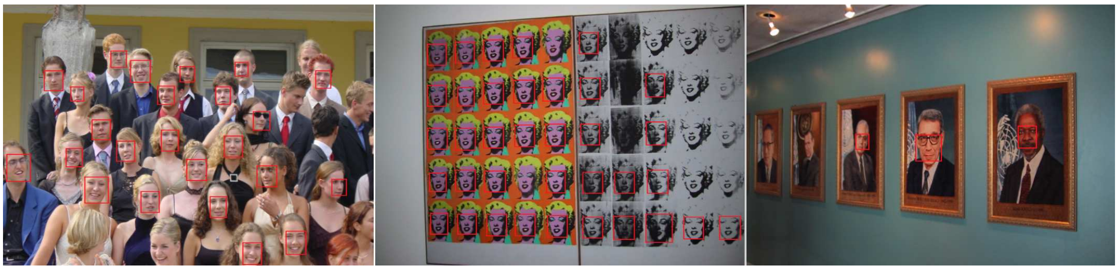


Figure 4.23: Examples of face detection results achieved by the classifier cascade trained using Asymmetric AdaBoost.

As suggested by Lienhart [158] and others, a number of improvements to the Viola-Jones approach are possible. In this chapter, an improved learning algorithm known as Asymmetric AdaBoost (AsymBoost, [282]) was implemented which is able to exploit the asymmetric requirement for very low false reject rate (typically $\leq 0.4\%$) relative to false positive rate (60% suffices for a cascade with 25 layers) at each cascade layer. As can be seen in figure 4.22, AsymBoost resulted in a small but significant improvement over AdaBoost in terms of the performance of the resulting classifier as evaluated using the large MIT+CMU frontal face image set. Figure 4.22 also shows the increasing complexity in terms of the number of "weak" component classifiers at each level of the cascade for the AsymBoost face detector. For example, the first 5 layers comprise a total of only 44 primitive bar features and yet are sufficient to reject 92.2% of non-faces. This new face detector was applied to the aforementioned image collection. Figure 4.23 shows some examples of face detections achieved by the algorithm. Compared to the method in section 4.5, execution time was greatly reduced, as was the number of false detections, although far fewer non-frontal faces are detected by the new method.

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).

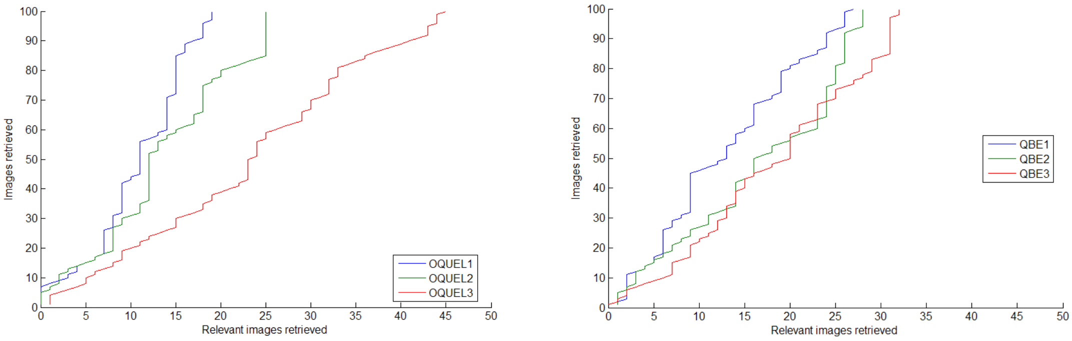


Figure 4.24: Plots of total number of images retrieved versus number of relevant images retrieved for *left*: OQUEL queries, *right*: query-by-example (QBE). In each case, results are shown for an initial query and two iterations of query refinement.

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 [222], 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 4.24 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. Search times were in the order of a few seconds for each query.

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 4.24. 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 capture the semantics behind the retrieval requirement adequately, even though all queries have access to the same set of image descriptors.

4.8 Conclusions

4.8.1 Discussion

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

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

4.8.2 Summary and Outlook

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

This chapter argues that query languages provide a flexible way of dealing with problems commonly encountered in CBIR, such as ambiguity of image content and user intention and the semantic gap which exists between user and system notions of relevance. By basing such a language on an extensible ontology, one can explicitly state ontological commitments about categories, objects, attributes, and relations without having to pre-define any particular method of query evaluation or image interpretation. A central theme of the chapter is the distinction between query description and image description languages, and the power of a formally specifiable language featuring syntax and semantics in order to capture meaning in images relative to a query. The combination of individually weak and ambiguous clues to determine object presence and estimate overall probability of relevance builds on recent approaches to robust object recognition and can be seen as an attempt at extending the success of indicative methods for content representation in the field of text retrieval [255, 220, 254, 299].

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

The use of more sophisticated natural language processing techniques would ease the current grammatical restrictions imposed by the syntax and allow statistical interpretation of more free-form query sentences consisting of words from an extended vocabulary. While this would also add an additional element of ambiguity, it would give users greater freedom to incorporate prior knowledge into the linguistic structure of their queries.

Moreover, ongoing efforts aim to acquire the weighting of the Bayesian inference nets used in scene and object recognition using a training corpus and prior probabilities for the visual categories. The goal is to reduce the need for pre-wired knowledge such as “an image containing regions of snow and ice is more likely to depict a winter scene”. An approach such as [70] paired with the structural Expectation Maximisation method might provide a means of automatically acquiring new high-level terms and their inference networks. The automated discovery of domain and general purpose ontologies together with the means of relating these to lower level evidence is an important challenge for data mining and machine learning research.

Chapter 5

Ontology-guided Dynamic Scene Understanding

5.1 Overview

This chapter describes an extension of the ontological inference framework to video and introduces mechanisms required to ground ontological descriptors given a set of visual processing modalities and a domain description.

It shows how modern techniques for structure and parameter learning in Bayesian networks can be applied to a labelled video data set to automatically generate effective high-level state and event recognition mechanisms for video analysis. Manual annotations consisting of high-level descriptors for actions and events in a surveillance scenario are combined with visual tracking and appearance modelling modules. Both the structure and parameters of Bayesian networks are then trained to infer high-level object and scenario labels on the basis of the visual properties and an ontology of states, roles, situations and scenarios which is derived from the original ground truth schema.

The integration of these different sources of evidence is optimised with reference to the syntactic and semantic constraints of the ontology. Through the application of these techniques to a visual surveillance problem, it is shown how high-level event, object and scenario properties may be inferred on the basis of the visual content descriptors and an ontology which is derived from the pre-defined ground truth schema. Performance analysis of the resulting framework allows alternative ontologies to be compared for their self-consistency and realisability in terms of the different visual detection and tracking modules.

5.2 Introduction

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 (known as "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.

As visual surveillance applications become increasingly prevalent, automated techniques for the detection and analysis of objects and events in video data are gaining prominence. It is likely that an increased reliance on such methods will bring about important changes to the way that research and development in relevant fields of computer vision is conducted and assessed. The case of vision-based biometrics in recent years offers some insights into likely developments in other areas of computer vision that are pertinent to the booming security industry.

Increased commercial and government interest in automated visual surveillance is not only resulting in increased emphasis on performance analysis and evaluation standards, but also fundamentally affects the way such research is conducted. Rather than focussing on the particular merits and intellectual importance of particular vision algorithms and representations, developers of visual surveillance systems will be confronted with largely externally imposed specifications of what information such systems are to extract from available video footage.

In this chapter, video sequences and ground truth from the CAVIAR project[1] 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

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

and appearance change which could easily be computed by means of a visual blob 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 that appear in the ground truth. From this one can draw conclusions as to the semantic and syntactic self-consistency and completeness of the ground truth schema, and the extent to which the manual annotations are consistent with the assumptions incorporated into the ontology. Such results then allow one to iteratively refine both the ontology and the underlying visual content extraction methods in order to arrive at a complete system that meets its requirements. They may also serve as a valuable basis for comparison of alternative approaches to solving a particular set of tasks.

5.3 Data Set

The aforementioned CAVIAR data comprises 28 sequences taken by a surveillance camera in the entrance lobby of the INRIA Rhone-Alpes research laboratory in Montbonnot, 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.

Each sequence has been manually annotated with the spatial location, angle of rotation, and extent of bounding boxes around individuals and groups of people.

Each such box is associated with a list of annotations (as will be explained in section 5.5.1) and a numerical label to identify it in subsequent frames. Annotations consist of binary states and probabilities (which are effectively either 1 or 0 in the ground truth) for the events, scenarios, situations and roles that are deemed to best describe the behavioural and situational context of the given person or group. Groups have a different set of descriptors from individuals and are defined in terms of their constituent individuals and the smallest bounding box that encompasses them.

However, this chapter limits its scope to the individuals and their descriptors. This is because high-level analysis on the basis of individual state and actions was found sufficient to investigate the use of the extended ground truth ontology for the creation and evaluation of Bayesian inference networks. Moreover, the criteria for grouping in the data were found to be somewhat ill-defined, for example often people are grouped together at a particular frame on the basis that they will interact in some way several seconds later in the sequence.

Training of Bayesian networks was carried out on approximately half the data, comprising a total of about 13000 video frames and over 17000 annotations of individual people, with the remainder reserved for evaluation purposes.

5.4 Visual Analysis and Tracking

This section provides an overview of the visual tracking and object modelling methods that were applied to the CAVIAR video sequences.

5.4.1 Background Modelling and Foreground Detection

The system maintains a background model and foreground motion history (obtained by frame differencing) that are adapted over time using an exponential rate of decay to determine the decreasing influence of previous values M_{t-k} in the history:

$$M_t = \alpha \; |\mathrm{im}_t - \mathrm{im}_{t-1}| + (1-\alpha) \; M_{t-1} \tag{5.1}$$

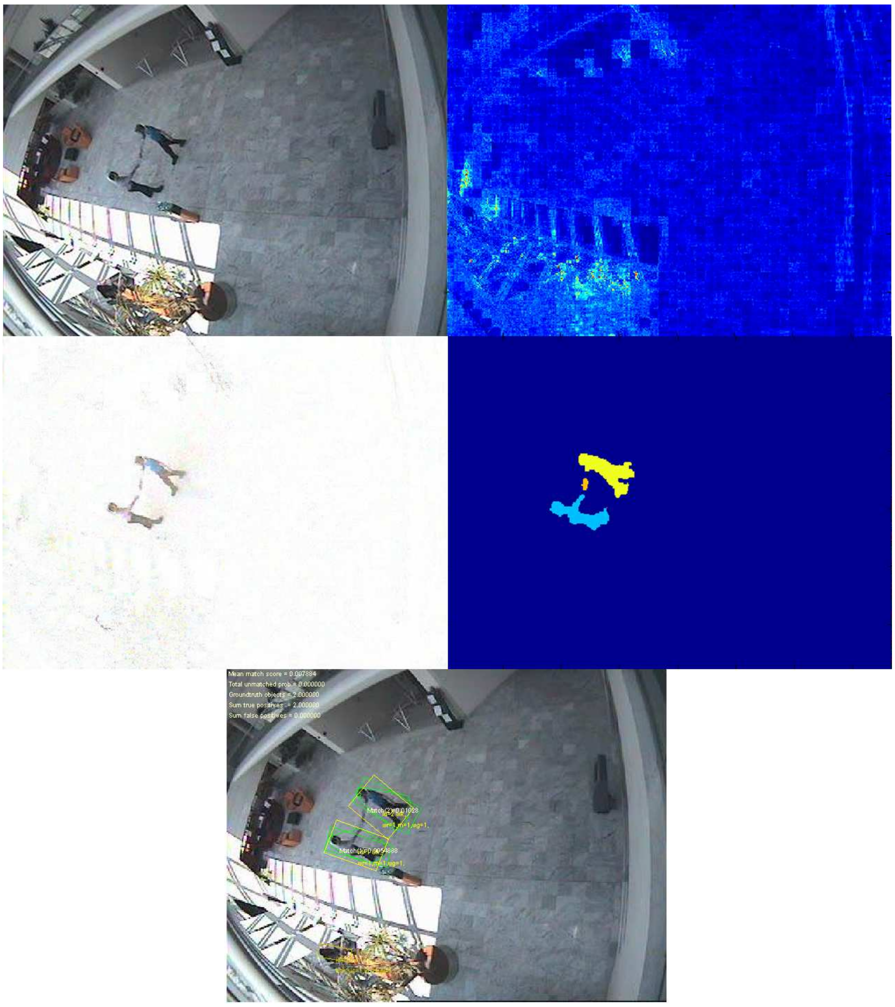

Mean match score = 0.007884
Total unmatched prob = 0.000000
Groundtruth objects = 2.000000
Sum true positives = 2.000000
Sum false positives = 0.000000
Match(2)=0.01028
wr=1,m=1,wg=1,

Figure 5.1: Foreground detection and tracking results (from left to right, top to bottom): 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 analysis results shown in yellow.

where im_t denotes the current image (video frame)[2], im_{t-1} is the frame at the previous time step, and $\alpha = 1 - e^{-1/\lambda_M}$. It was found sufficient to set λ_M to some constant value such as 3.

[2] Unless noted otherwise, computations involving M_t, im_t, and B_t are implemented by maintaining separate matrices for the red, green, and blue (RGB) colour channels. The operation $||M||$ denotes summation over the RGB matrix elements to arrive at a single image intensity value (scaled to lie within $[0,1]$) at each pixel location $M(x,y)$.

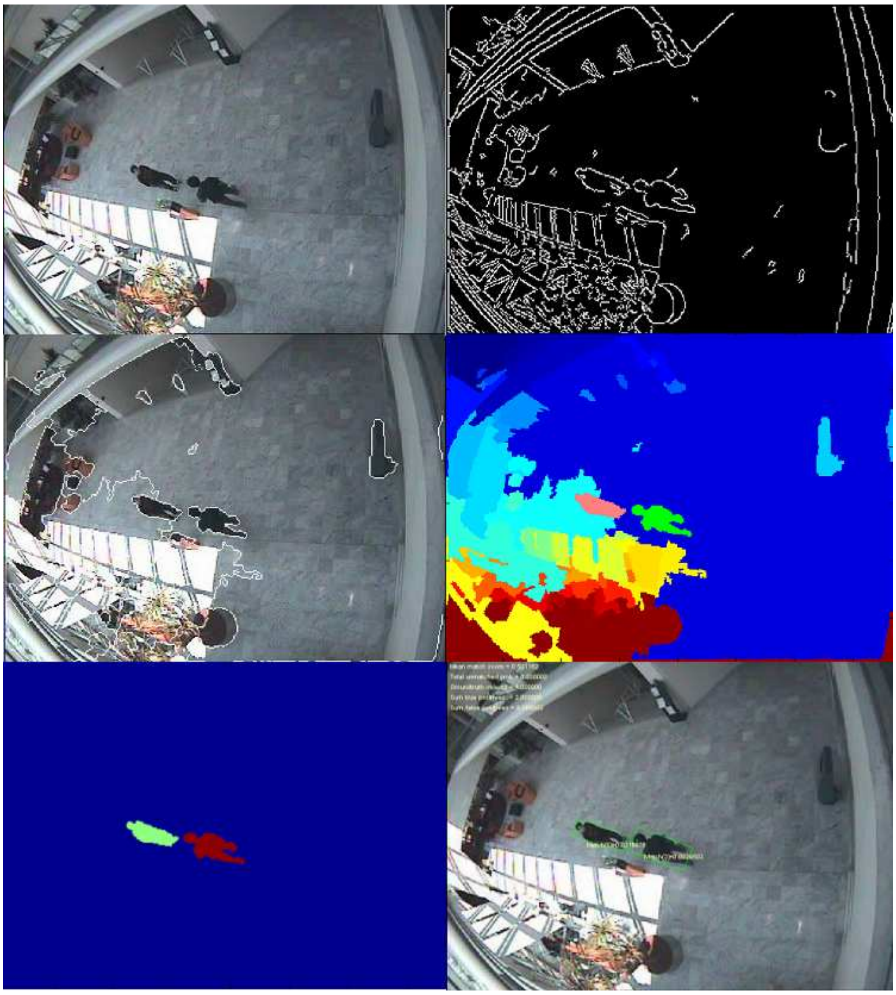


Figure 5.2: Foreground detection and tracking with background segmentation (from left to right, top to bottom): Original frame; Detected edges; Boundaries of segmented regions; Pixel membership of segmented regions; Regions identified as foreground; Resulting tracked objects (outlined in green) with ground truth data and analysis results shown in yellow.

The motion history M_t at time t is used to identify a background image mask bim_t of pixels undergoing sufficiently slow change (i.e. due to noise or gradual changes in lighting conditions) which can then be used to reliably update the background model B_t and estimate its variance:

$$B_t = \neg\,\mathrm{bim}_t\,.B_{t-1} \;+\; \mathrm{bim}_t\,.\,[\beta\ \mathrm{im}_t + (1-\beta)\ B_{t-1}]\,;\quad B_0 = \mathrm{im}_0 \tag{5.2}$$

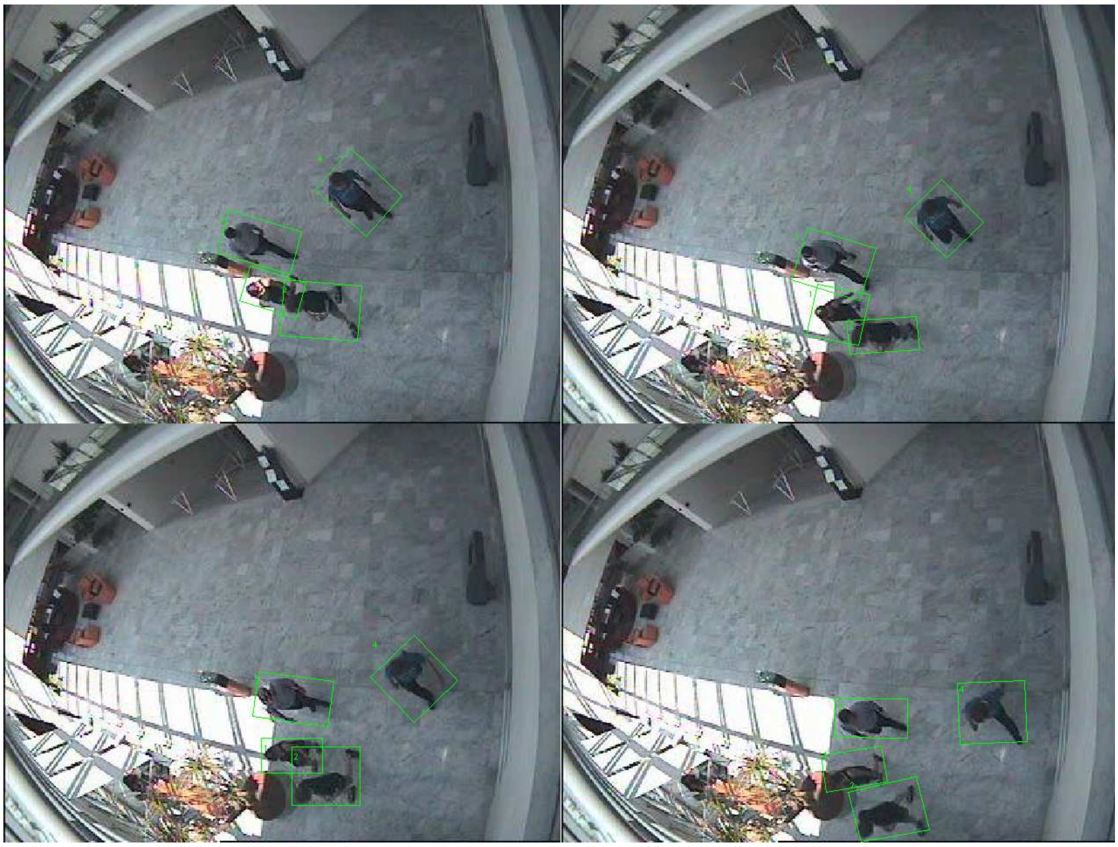


Figure 5.3: Sample frames from a surveillance sequence showing the objects tracked by the blob tracking framework outlined in green.

where "." denotes the element-wise (Hadamard) product of two matrices, $\beta = 1 - e^{-1/\lambda_B}$, and

$$\mathrm{bim}_t(x, y) = \begin{cases} 1 & \text{if } ||M_t(x, y)|| < \tau \\ 0 & \text{otherwise} \end{cases} \tag{5.3}$$

is a binary image mask identifying those pixels (x, y) in the motion history image M_t whose intensity is below a threshold τ, and $\neg\,\mathrm{bim}_t$ is the logical inverse of bim_t. Based on empirical performance over the training data, values of $\lambda_B = 3$ and $\tau = 0.12$ were chosen.

In order to be considered as part of the non-static foreground, pixels must first be identified as outliers of the background process. Outliers are those that exceed a difference threshold which is a multiple of the estimated background variance, σ_t^B, relative to B_t. A pixel at image location (x, y) is considered an outlier if

$B_t^{\text{outlier}}(x,y) = 1$, where B_t^{outlier} is a binary mask defined by

$$B_t^{\text{outlier}}(x,y) = \begin{cases} 1 & \text{if } \big|\, \|\text{im}_t(x,y)\| - \|B_{t-1}(x,y)\| \,\big| > k\ \sigma_t^B(x,y) \\ 0 & \text{otherwise} \end{cases} \tag{5.4}$$

The background variation at each pixel is estimated using

$$\sigma_t^B = \neg\,\text{bim}_t\,.\sigma_{t-1}^B \ + \ \text{bim}_t\,.\left[\gamma\ \big|\, \|\text{im}_t\| - \|\text{im}_{t-1}\| \,\big|\ + (1-\gamma)\ \sigma_{t-1}^B\right] \tag{5.5}$$

In practice, setting $k = 5$ and $\gamma = 1 - e^{-1/5}$ yielded an acceptable trade-off of false positives and false negatives.

Pixels which are thus classified as outliers B_t^{outlier} with respect to the background are only labelled as being part of the foreground F_t if they are not likely to be part of (moving) shadows as determined by the DNM1 (deterministic non-model based) algorithm described in [211]:

$$F_t(x,y) = B_t^{\text{outlier}}(x,y) \wedge \neg\,\text{shadow}_t(x,y) \tag{5.6}$$

The shadow detection algorithm is based on the observation that pixels which have become part of a shadow retain similar hue (H) and saturation (S) levels but exhibit reduced luminance (V). In this chapter, the background image B_t is used as the basis for comparison since it provides a robust representation of pixel colour acquired over a longer time scale:

$$\text{shadow}_t(x,y) = \begin{cases} 1 & \text{if } \left(c_{V1} \leq \frac{V_{\text{im}_t}(x,y)}{V_{B_t}(x,y)} \leq c_{V2}\right) \\ & \wedge\,(\Delta H_{\text{im}_t,B_t} \leq c_H) \wedge (\Delta S_{\text{im}_t,B_t} \leq c_S) \\ 0 & \text{otherwise} \end{cases} \tag{5.7}$$

where c_{V1}, c_{V2}, c_H, and c_S are constants chosen to reflect the lighting conditions and noise levels in particular parts of the scene.

5.4.2 Blob Tracking and Adaptive Appearance Modelling

Foreground pixels are clustered using connected component analysis to identify moving regions ("blobs"). In order to improve the spatial localisation of blobs, the

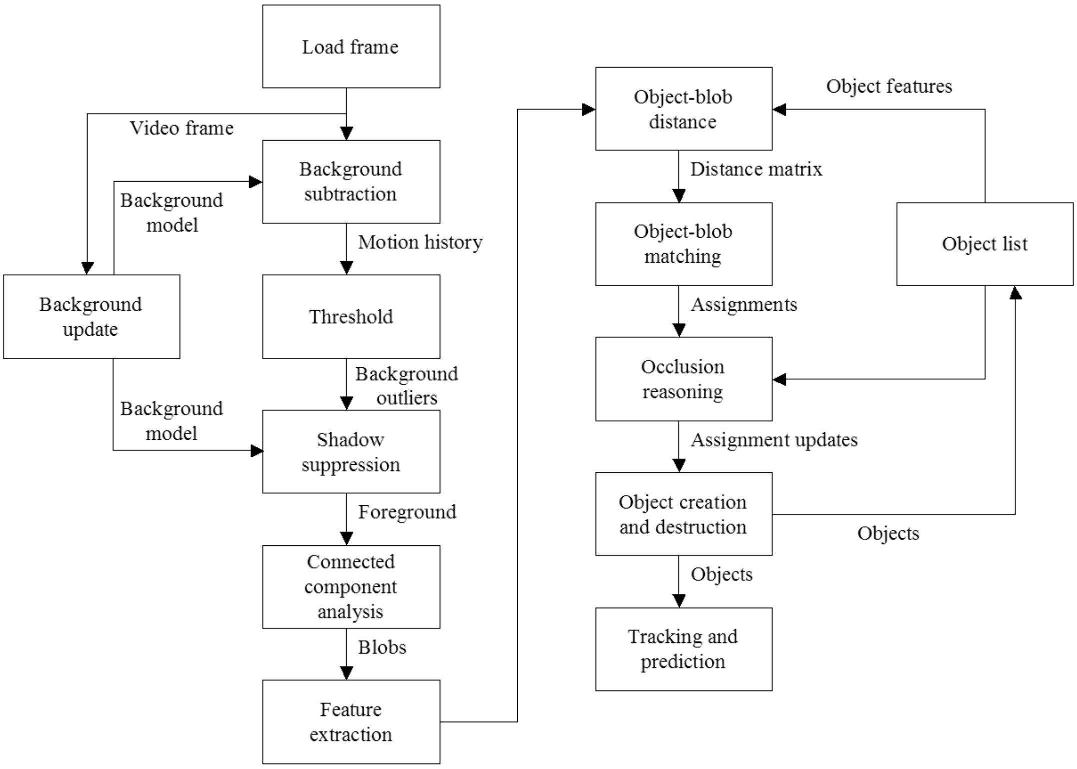


Figure 5.4: Overview of the blob tracking method.

static image segmentation method described in section 4.5.2 is applied to individual video frames. Segmented regions which significantly overlap with candidate foreground regions and whose constituent pixels differ from the current background by more than a threshold are identified as foreground objects. This reduces problems such as fragmentation of foreground objects into multiple blobs. It could also be used to provide tracking and motion analysis at finer granularity, e.g. for gesture or gait recognition, or to detect activities such as hand shakes.

Detected blobs are then parameterised using shape (bounding box, centre of gravity, major axis orientation) and colour measures. Colour appearance is modelled by means of both an RGB histogram and a Gaussian mixture model in hue-saturation space. In a similar manner to [171], re-estimation of the mixture parameters is performed selectively by weighting frame contributions with the blob's colour log-likelihood under the model.

Blob positions are tracked using a Kalman filter or Condensation tracker 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 shape and appearance. Histograms are compared using the EMD (Earth Mover's Distance, [229]) measure and provide a useful measure of short-term appearance variation, while the Gaussian mixture models a more stable and long-term representation of appearance that is useful for identity maintenance across object occlusions.

Figure 5.4 summarises the blob tracking framework. Figure 5.1 shows results from the background modelling, foreground detection, and visual tracking for one frame of the CAVIAR sequences. As shown in figure 5.2, objects can often be localised more accurately in practise by integrating information from the image segmentation.

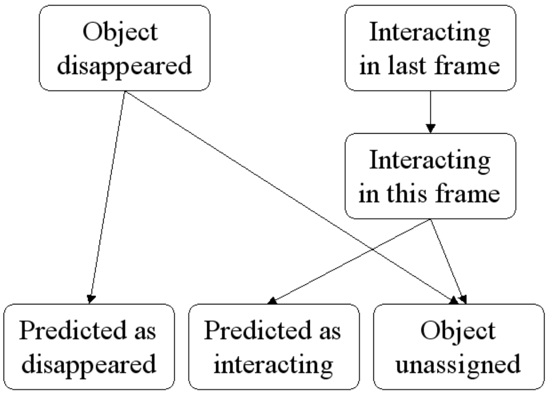


Figure 5.5: Bayesian network for occlusion reasoning and prediction of object interactions.

5.4.3 Occlusion Reasoning

To make tracking more robust, the object to blob assignment stage features a Bayesian network for reasoning about occlusions and object interactions (see figure 5.5) based on observed or predicted overlap of object bounding boxes and failures of object assignment.

Together with the blob appearance models discussed in section 5.4.2, this allows separate object identity to be maintained across partial or total occlusions.

If two or more objects overlap, they may initially be inscribed within a single blob. However, by matching their individual appearance models to their predicted locations in the current frame, one can separate overlapping objects or determine which object is in the foreground. Figure 5.3 shows results of vision-based tracking over several frames of one of the CAVIAR sequences.

5.4.4 Visual Tracking Accuracy

A range of performance evaluation metrics have been proposed to assess the quality of visual object tracking [74, 22]. Important factors include the number of:

- *True positives* N_{tp} ("hits"): the number of visually tracked objects confirmed by the ground truth
- *False positives* N_{fp} ("duds"): the number of objects not matching the ground truth
- *False negatives* N_{fn} ("misses", "false rejects"): the number of ground truth objects not matched by the tracking method
- *True negatives* N_{tn} ("true rejects")[3]: the number of erroneous observations rejected by the visual tracking

It follows that the number of objects in the ground truth, N, can be expressed as $N = N_{\mathrm{tp}} + N_{\mathrm{fn}}$. One can then define metrics such as

- *Detection rate* (sensitivity)

$$\mathrm{DR} = \begin{cases} \frac{N_{\mathrm{tp}}}{N} & \text{if } N > 0 \\ 1 & \text{otherwise} \end{cases}$$

- *False positive rate*

$$\mathrm{FR} = \begin{cases} \frac{N_{\mathrm{fp}}}{N_{\mathrm{fp}}+N} & \text{if } N_{\mathrm{fp}} + N > 0 \\ 0 & \text{otherwise} \end{cases}$$

[3] This measure is less useful in practice since the number of potential observations will differ significantly between tracking systems and is likely to be very large.

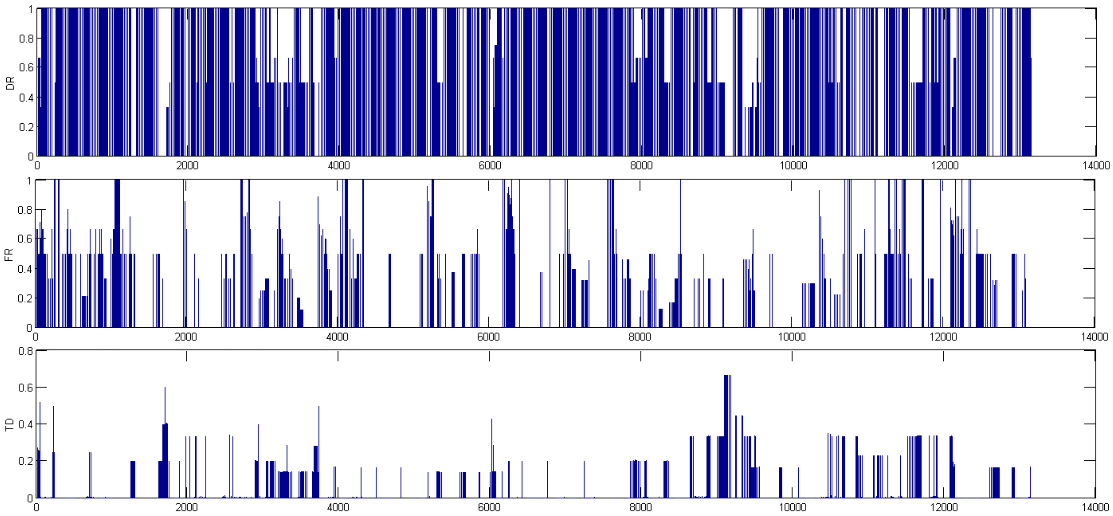


Figure 5.6: Performance results of the visual tracking for about 13000 frames from the CAVIAR sequences. ***Top*****: detection rate** DR**.** ***Middle*****: false positive rate** FR**.** ***Bottom*****: Average distance from track** TD**.**

- *Recall* = mean(DR)
- *Precision* = $\text{mean}(\frac{N_{\text{tp}}}{N_{\text{tp}}+N_{\text{fp}}})$

A mean distance-from-track measure TD is also computed. For each object in the ground truth, this is the average distance across the sequence of the normalised (in terms of the maximum possible distance across the image) Euclidean distance of object and matching observation centres of gravity. For ground truth objects not tracked, TD is set to the maximum value of 1.

Figure 5.6 shows results of DR, FR and TD calculated for the visual object tracking framework presented above over each of the subset of approximately 13000 frames from the CAVIAR sequences used for evaluation. The overall mean detection rate, i.e. recall, is 0.84 at an average precision of 0.52. The mean false alarm rate is 0.23 with mean TD of 0.06.

The detection rate is usually lower at the start of a sequence, largely because there may already be annotated objects present at the beginning that have not yet moved and thus have not yet been detected by the tracker. The false alarm rate is a relative measure of how many false alarms there are relative to the actual number of ground truth objects. A value of 0 thus means that there are no false positives, and FR approaches 1 as the number of false positives grows large or if there are no ground truth objects present in a given frame. Many of the reported false positives for the CAVIAR sequences are due to people moving in the sequence who are not part of the ground truth. This is because the sequences often feature both actors, whose position and actions have been annotated as part of the CAVIAR project, and occasional bystanders, who do not feature in the annotations but are nevertheless correctly identified and tracked by the system presented in this chapter. For example, certain areas visible in each frame, such as the reception desk (lower left) and first floor corridor (upper left), were not considered by the annotators and consequently people moving in these areas do not feature in the ground truth. If the aforementioned regions are excluded from consideration by the tracking methods, the average precision relative to the ground truth annotations increases substantially.

5.5 High-level Analysis

5.5.1 Domain Ontology

Ontologies encode the relational structure of concepts which one can use to describe and reason about aspects of the world. An ontology must however be grounded in reality, i.e. the data it is to process. This may be achieved either by hierarchically re-expressing higher terms of the ontology in terms of well-defined primitives, or by providing a sufficient number of examples of the desired concepts such that the system may internalise them by means of machine learning techniques.

The CAVIAR annotations can naturally be organised into a hierarchical ontology as shown in table 5.1. 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 broad hierarchical relationships can be used as a structural prior for the training of Bayesian networks as described in the next section.

In order to study the extent to which elements of the ontology may be inferred on the basis of automatically extracted visual information, one needs to augment

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

Table 5.1: CAVIAR ontology.

the ontology with appropriate descriptors that can be computed from raw sequence data using computer vision techniques. Using the tracking and appearance modelling framework described in section 5.4, a set of such descriptors was defined in order to form the bottom layer of the ontology:

- *cvSpeed*: Current object speed (as estimated from the tracker) in terms of the estimated displacement of the object's bounding box expressed in pixels per second (calculated per frame and normalised using the camera's frame rate)

- *cvFlow*: Amortised flow measure representing a recent history of the object's motion:

$$\mathrm{fl}_t = \gamma\left(\left|\mathrm{cx}_t - \mathrm{cx}_{t-1}\right| + \left|\mathrm{cy}_t - \mathrm{cy}_{t-1}\right|\right) + (1-\gamma)\,\mathrm{fl}_{t-1}$$

where $\gamma = 1 - e^{-1/3}$ and $(\mathrm{cx}_t, \mathrm{cy}_t)$ = object centre of gravity at time t.

- *cvLifetime*: Whether or not the object has been newly instantiated or is about to be terminated due to no longer being detectable in the image (in the absence of any other explanation offered by the occlusion reasoning).

- *cvHistdist*: Measure of inter-frame appearance variation calculated as the weighted sum of histogram EMD measure and Gaussian mixture model likelihood. The weight given to histogram distance is increased if the object is moving rapidly.

- *cvOccstat*: Whether the object is estimated to be unoccluded, occluded, or to have disappeared based on the occlusion reasoning network shown in figure 5.5.

These visual descriptors 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 5.4, and offer a reasonable basis for studying the requirements for low-level analysis mechanisms that result from the pre-defined ontology of higher-level terms. Additional object tracking and analysis modules could easily be integrated into the existing framework to provide additional information for terms which are currently hard or impossible to infer (e.g. the detection of abandoned objects and other roles that require knowledge of multi-object interactions).

The principle goal of this investigation was to study the suitability of the ontology and ground truth for automated construction of Bayesian inference networks independent of the performance of any particular tracking methods. Thus the available annotations were used to initialise objects maintained by the visual tracking framework in order to then augment the ground truth for each individual with the resulting visual descriptors listed above.

5.5.2 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 [107, 141] for an overview and further references. In this chapter, 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 training set (augmented as required with the information gathered by the computer vision techniques). All nodes were represented as discrete states, with the nodes cvSpeed, cvFlow and cvHistdist quantised to 5 different values. Nodes cvLifetime and cvOccstat have 3 states while all other variables are binary.

Since the nodes are discrete, learning the parameters of the conditional probability density for a node X_i requires one to learn the entries of a discrete conditional probability table θ_{ijk}, which specifies the probabilities of the node assuming each of its possible values k for each combination of value j of its parent nodes $\mathrm{Pa}(X_i)$:

$$\theta_{ijk} = P(X_i = k | \, \mathrm{Pa}(X_i) = j); \quad \sum_k \theta_{ijk} = 1 \;\; \forall i, j \tag{5.8}$$

The structure of a Bayesian network consists of a directed acyclic graph (DAG) G whose connectivity matrix defines the conditional (in)dependence relationships among its constituent nodes X and hence defines the form of the conditional probability tables.

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 [182]:

- The K2 algorithm [51] 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 [84], 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,

$$P(G|D) = P(D|G)P(G) \tag{5.9}$$

  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

$$P(D|G) = \int_{\theta} P(D|G, \theta)P(\theta|G) \tag{5.10}$$

  where D is the training data, G is the graph structure, and θ are the network parameters. Assuming parameter independence, the marginal likelihood can then be computed efficiently by decomposing it into a product of terms over the N nodes in the network (see [183]):

$$P(D|G) = \prod_{i=1}^{N} \int_{\theta_i} P(X_i|\operatorname{Pa}(X_i), \theta_i)P(\theta_i) \tag{5.11}$$

  In the case of discrete nodes, the parameters θ_i are multinomial distributions and the integrals can be computed analytically.

- The Bayesian Information Criterion (BIC), which approximates the marginal likelihood using a Minimum Description Length (MDL) approach. Following

[108], 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:

$$\log P(D|G) \approx \log P(D|G, \hat{\theta}_G) - \frac{d}{2}\log M = \mathrm{BIC}(D, G) \tag{5.12}$$

where M is the number of training cases in D, $\hat{\theta}_G$ is the maximum likelihood estimate of the parameters (see equation 5.16 below), and d is the number of free parameters (degrees of freedom) of the model. The BIC score can also be computed efficiently by decomposing it into a product of local terms for each node in the network. In the present case, assuming multinomial parameter distributions allows one to write [183]:

$$\begin{aligned} \mathrm{BIC}(D, G) &= \sum_i \left[\sum_m \log P(X_i |\, \mathrm{Pa}(X_i), \hat{\theta}_i, D_m) - \frac{d_i}{2}\log M \right] \\ &= \sum_i \left[\sum_{jk} N_{ijk} \log \theta_{ijk} - \frac{d_i}{2}\log M \right] \end{aligned} \tag{5.13}$$

where d_i is the number of parameters in the conditional probability table associated with node X_i and the other variables are as defined in equation 5.15 below.

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 using maximum likelihood estimation using the log-likelihood of the training data. Assuming independence of the training cases, the log-likelihood of the training set $D = \{D_1, \ldots, D_M\}$ can be computed over all N nodes of the network as follows:

$$\mathrm{LL} = \log \prod_{m=1}^{M} P(D_m|G) = \sum_{i=1}^{N}\sum_{m=1}^{M} \log P(X_i |\, \mathrm{Pa}(X_i), D_m) \tag{5.14}$$

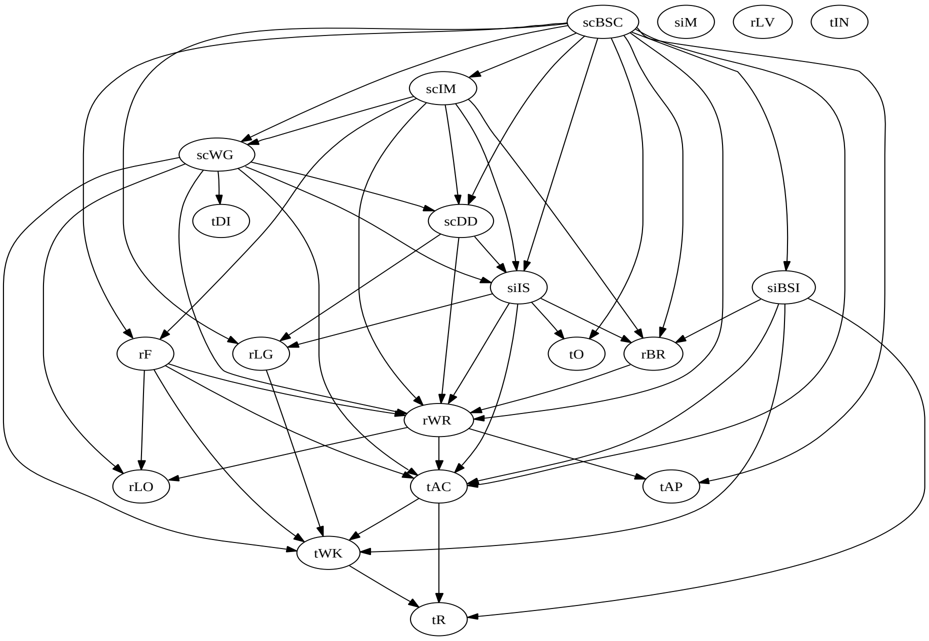


Figure 5.7: Bayesian network structure trained using the K2 algorithm applied to the original ground truth schema.

where $\mathrm{Pa}(X_i)$ are the parents of node X_i. As described in [183], in the case of

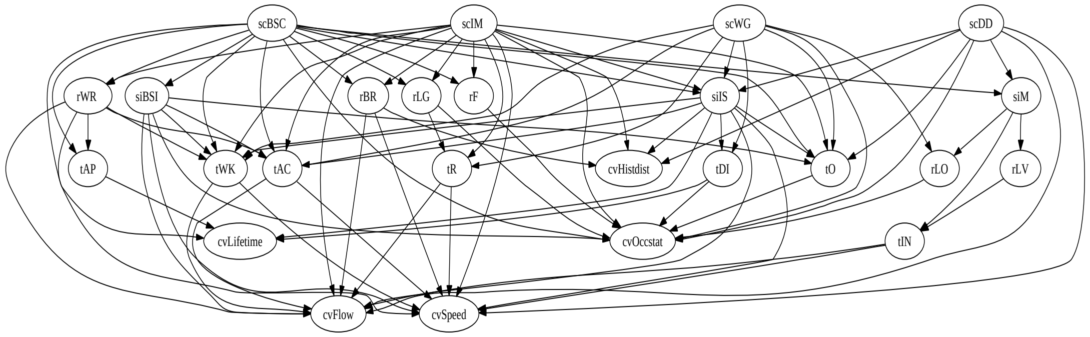


Figure 5.8: Bayesian network structure trained using the K2 algorithm with a structural prior and using the extended ontology.

discrete nodes with tabular conditional densities θ_{ijk}, this can be re-written as

$$\begin{aligned}\mathrm{LL} &= \sum_i \sum_m \log \prod_{j,k} \theta_{ijk}^{I_{ijkm}} \\ &= \sum_i \sum_m \sum_{j,k} I_{ijkm} \log \theta_{ijk} \\ &= \sum_{ijk} N_{ijk} \log \theta_{ijk} \qquad (5.15)\end{aligned}$$

where the indicator function $I_{ijkm} = I(X_i = k, \mathrm{Pa}(X_i) = j | D_m)$ takes the value 1 if the event $(X_i = k, \mathrm{Pa}(X_i) = j)$ occurs in case D_m (and $I_{ijkm} = 0$ otherwise), and hence $N_{ijk} = \sum_m I(X_i = k, \mathrm{Pa}(X_i) = j|D_m)$ is simply a count of the number of times that $(X_i = k, \mathrm{Pa}(X_i) = j)$ occurs in the data. From this one can derive the maximum likelihood estimate (MLE) of the parameters:

$$\hat{\theta}_{ijk} = \frac{N_{ijk}}{\sum_{k'} N_{ijk'}} \qquad (5.16)$$

5.5.3 Performance Analysis

Figure 5.7 shows the graph structure which results from training a Bayesian network using the K2 method and the data from the CAVIAR ontology in table 5.1. As can be seen from the graph, several of the nodes are not definable in terms of the other nodes and the overall connectivity seems somewhat ad-hoc. By contrast, figure 5.8 shows a network that was trained using the full ontology 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.

In order to arrive at a simple statistic for comparison of different networks, one can compute the average likelihood of the data per node:

$$L^{\sim} = e^{\mathrm{LL}/(MN)} \qquad (5.17)$$

Using this measure, the network in figure 5.7 achieves a score of $L^{\sim} = 0.729$, while the network in figure 5.8 scores $L^{\sim} = 0.888$. Its likelihood scores for each node are shown in table 5.2.

Node	Order	Likelihood	Node	Order	Likelihood
scBSC	1	0.604	tAP	14	0.959
scIM	2	0.534	tDI	15	0.980
scWG	3	0.505	tO	16	0.978
scDD	4	0.867	tIN	17	0.999
siM	5	0.999	tAC	18	0.728
siIS	6	0.832	tWK	19	0.718
siBSI	7	0.902	tR	20	0.937
rF	8	0.904	cvSpeed	21	0.266
rBR	9	0.845	cvFlow	22	0.278
rLV	10	0.999	cvLifetime	23	0.993
rLG	11	0.927	cvHistdist	24	0.392
rWR	12	0.695	cvOccstat	25	0.555
rLO	13	0.985			

Table 5.2: Likelihood scores for the terms in the augmented ontology as computed for the Bayesian network shown in figure 5.8. The topological ordering of the nodes used by the K2 learning algorithm is also indicated.

These scores indicate how well the conditional distribution at each node represents the training data for that node. For example, the values derived using computer vision have no support in the ground truth and their derivation therefore involves a greater amount of uncertainty, resulting in a lower likelihood score. Even within the nodes corresponding to terms in the ground truth, some such as “rWR”, “scWG”, “scIM” and “scBSC” exhibit smaller likelihood values. This may be due to inconsistencies in the ground truth or failure of the Bayesian network to fully model their interdependencies.

Perhaps more useful conclusions can be drawn from an analysis of posterior probability scores. Using a suitable inference algorithm such as the Junction tree method, one can compute the marginal probabilities for each possible state of a given node for the available evidence. By computing the marginal probability of a given node for its “correct state” (i.e. that provided in the ground truth data), one can compute the expected detection rate for a given term in the ontology as

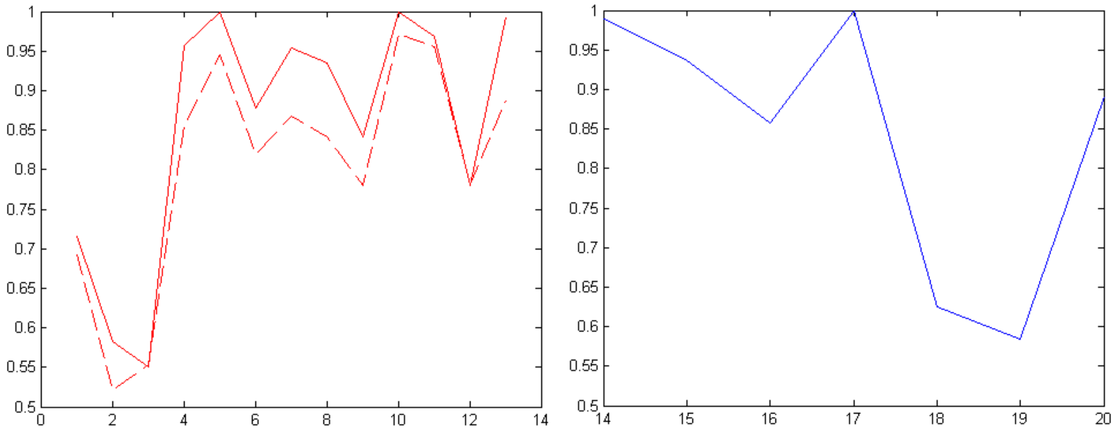


Figure 5.9: Expected detection rates for the nodes in the Bayesian network in figure 5.8 given different evidence (see section 5.5.3). The values on the x-axis correspond to nodes in the network when enumerated in topological order as in table 5.2 (i.e. scBSC=1, scIM=2, ..., cvOccstat=25).

the mean of these marginal scores over the evaluation data set. This allows one to quantify the value of adding and removing nodes and edges in the Bayesian network.

The dashed red line on the left of figure 5.9 indicates how well defined the upper-level terms are given only the value of the states (tAP, tDI, tO, tIN, tAC, tWK, tR) as evidence for the network in figure 5.8. By comparison, the solid red line shows expected correct detection rates if the Bayesian network is given both the values of the states in the ground truth and those of the computer vision derived nodes as evidence. It can be seen that adding the latter improves the performance of higher-level inference. The blue line on the right side of figure 5.9 indicates the expected correct inferences for each state given the values of the computer vision nodes only.

5.6 Enhanced Visual Analysis

While the preceding sections demonstrate that the surveillance ontology can be grounded quite effectively using simple blob tracking methods, more sophisticated visual analysis is often desirable to improve recognition performance. This section presents some work towards this goal by introducing novel methods for combining contour, edge, and blob-based approaches to object tracking. These form the basis of an extended ontology of visual descriptors as shown in section 5.7.

5.6.1 Overview and Background

Most modern approaches to object tracking rely on parametric or non-parametric representations such as edges, active contours, colour blobs, or segmented regions. This section presents a framework that integrates information from these approaches in order to reliably track objects in sequences taken by a single static camera and ascribe parametric motion models to such objects. The resulting system combines many of the benefits of the individual methods such as self-initialisation, good convergence to object boundaries, and derivation of motion and appearance models that may serve as a basis for high-level activity analysis.

Following [177], most efforts in visual tracking can be described in terms of the underlying representation they employ:

- *Image-based* methods rely on tracking of features or spatio-temporal aspects derived directly from images.
- *Object-based* approaches perform a figure-ground segmentation in order to identify and track point groups, ellipses, bounding rectangles, blobs or regions.

The growing demand for high-level analysis of motion and behaviour in visual surveillance and other applications such as video editing and special effects have placed increased emphasis on the description and segmentation of motions in video. Techniques for motion estimation in visual tracking have taken varied approaches, which can be broadly grouped into two categories:

- *Optical-flow* techniques build a dense velocity vector-field in the image according to measured intensity changes at pixels in the image. Moving objects can then be identified as groups of pixels with the same motion.
- *Feature-based* techniques extract local descriptors such as edges or corners from the images and identify the corresponding features in subsequent frames.

As ever, different methods offer different advantages and disadvantages (see [275] for a discussion). Corner features (where the image changes in two directions) can be poor to track since they are small and sparse. This means they rarely supply enough information to group other pixels obeying the same motion, and also that they can easily be missing in subsequent frames due to occlusions and noise. In contrast, edges, which correspond to larger-scale structural attributes of objects, can provide information about the motion of many pixels, and are also more likely to be stable over short time scales. Unlike pixel-based techniques, one need not assume that the image intensity of each observed point in the world remains nearly constant. Most importantly, edges in the image generally correspond to some real-world contour such as outlines of objects and are thus a natural intermediary between feature and object based approaches to visual tracking and analysis.

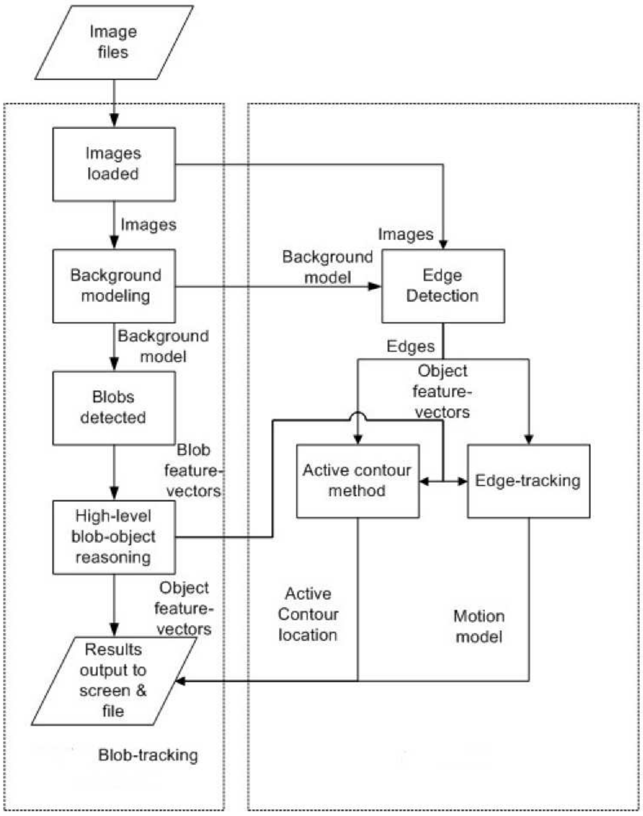


Figure 5.10: Overview of the combined contour, edge, and blob tracking framework.

This section presents work which integrates several state of the art approaches in object-based tracking and feature-based motion estimation. The resulting framework robustly identifies and parameterises the spatial extent and boundary of objects, estimates semi-parametric models of their colour appearance, and yields a parametric description of their motion.

As shown in figure 5.10, the system processes images (frames) from a video sequence. It maintains an adaptive background model in order to initially identify and group foreground pixels into coloured "blobs" (see 5.4) whose appearance is modelled using colour histograms and Gaussian Mixture models. The quality of the generated blobs can be further improved by means of a static edge detection and region segmentation as described in section 5.4.2. In order to parameterise object motions and deformations, a sample-based edge tracking method due to Smith [251] has been adapted. The method samples points along edges, tracks them over subsequent frames, and estimates a parameterised motion model for the whole edge from the observed motion of constituent pixels (see section 5.6.2).

In order to model and track the shape of objects in terms of their closed boundaries, approaches based on active contours ("snakes", [23]) have been very prominent in the vision community. In this thesis, the Gradient Vector flow method of Xu and Prince [297] was adapted in order to perform the transition from tracked edges to closed boundary contours (see section 5.6.3). The snake's external force is computed as a diffusion of the gradient vectors of an edge map derived using the edge detector.

Section 5.6.4 explains the approach to combining these methods in an integrated object tracking and analysis framework. Results for the combined approach are shown in section 5.6.6.

5.6.2 Sample-based Edge Tracking

In [251], an edge-based system for motion segmentation based on tracking sample points ("edgels") across a small number of frames in a sequence is presented. Here, this approach is adapted to the problem of object-tracking. Figure 5.11 gives an overview of the process.

At the beginning of the process for generating a motion model based on the movement between the first and second frames, the model is initialised as having

- In the first frame:
  - Compute edges using the segmentation algorithm.
  - Initialise the motion model for each edge.
  - Sample edgels k along each edge (for large objects, typically every 5th pixel is sampled) and calculate their colour image gradients using convolution.
- In subsequent frames:
  - For each edgel k in each edge:
    - Transform k according to the current motion model
    - Compute the unit normal to the edge $\hat{\mathbf{n}}^{(k)}$ at k
    - Search along the closest compass direction to the normal for a match in order to find the residual error, that is, the difference between the current motion model's positioning of the edge and the measured location.
  - Calculate an improved motion model for each edge by minimising the residual error. Iterate over the steps for the frame, improving the motion model until it converges.

Figure 5.11: Overview of the edge-tracking method

zero motion since motion in all directions is equally likely. If an edge is encountered along the search path in subsequent frames, its colour gradient is compared with that of the sample edgel in the previous frame to see if it matches. The pixel with the smallest change in colour gradient (as determined using a convolution mask) is taken as the matching pixel, provided it is below some threshold $\mathrm{diff}_{\mathrm{max}}$.

The motion model parameterises the observed transformations $\mathbf{z}$ of edge sample points across the sequence using a Lie group formalism. This consists of a weighted sum of deformations described by Lie vectors $\mathbf{L}_i$ weighted by parameters α_i such that $\mathbf{z}' = \mathbf{z} + \sum_i \alpha_i \mathbf{L}_i$. Following standard convention, the eight 2D projective deformations are x- and y-translation, rotation, dilation, pure shear, shear at 45^o, and deformation with finite vanishing point along x or y.

Estimation of the motion model of a given edge in terms of weights α_i is carried out by minimising the normal distances between the edgel locations predicted by the model and their actual locations in the current frame. For a given edgel k, the residual error $r^{(k)}$ is thus given by the difference between the current motion and that predicted by the current motion model parameters. The former is measured by the normal distance, $d^{(k)}$, of the edge at k in the current frame compared to the previous frame, and the latter by projecting the motion model onto the unit normal $\hat{\mathbf{n}}^{(k)}$ to the edge at point k:

$$r^{(k)} \;=\; d^{(k)} - \sum_i \alpha_i \left(\mathbf{L}_i^{(k)} \cdot \hat{\mathbf{n}}^{(k)} \right) \tag{5.18}$$

As pointed out in [251], the residual errors are unlikely to be normally distributed and in fact resemble a Laplacian distribution. To account for this, M-estimators are used, which allow minimisation of an arbitrary function of the residuals, $\rho(r^{(k)})$. An M-estimator that minimises $\rho(r^{(k)})$ is the maximum likelihood estimator for a set of residual errors $r^{(k)}$ with a probability distribution $P(r^{(k)}) \propto e^{-\rho(r^{(k)})}$, which allows one to choose a $\rho(r^{(k)})$ based on the error distribution. This distribution is assumed to be Gaussian for small values of $r^{(k)}$ and Laplacian otherwise.

In order to minimise $\rho(r^{(k)})$, one must evaluate $\sum\limits_{k=1}^{K} \frac{\partial \rho(r^{(k)})}{\partial \alpha_i} = 0$ for all deformation modes $i = 1...n$. By defining the *weight function*

$$w(r^{(k)}) = \frac{1}{r^{(k)}} \frac{d\rho(r^{(k)})}{dr^{(k)}} \tag{5.19}$$

one can re-write this as

$$\sum_{k=1}^{K} \frac{\partial \rho(r^{(k)})}{\partial \alpha_i} = \sum_{k=1}^{K} \frac{d\rho(r^{(k)})}{dr^{(k)}} \frac{\partial r^{(k)}}{\partial \alpha_i} = \sum_{k=1}^{K} w(r^{(k)}) r^{(k)} \frac{\partial r^{(k)}}{\partial \alpha_i} = 0 \qquad \text{for } i = 1...n \tag{5.20}$$

which leads to the same system of equations as the *weighted least squares* problem, where one is required to minimise $\sum\limits_{k=1}^{K} w(r^{(k)})\, {r^{(k)}}^2$ for the vector of parameters $\boldsymbol{\alpha}$. This is solved by an iterative process in which the weights are updated at each iteration until convergence, using the errors $\mathbf{d}$ produced by the current parameters of the motion model.

One can now solve $\boldsymbol{\alpha} = \mathbf{M}^{-1}\mathbf{v}$ with

$$\mathbf{v} = \mathbf{N^T} \begin{pmatrix} w^1 d^1 \\ w^2 d^2 \\ \vdots \\ w^{(k)} d^{(k)} \end{pmatrix} \; ; \quad \mathbf{N} = \begin{pmatrix} \mathbf{L}_1^1 \cdot \hat{\mathbf{n}}^1 & \mathbf{L}_2^1 \cdot \hat{\mathbf{n}}^1 & \dots & \mathbf{L}_n^1 \cdot \hat{\mathbf{n}}^1 \\ \mathbf{L}_1^2 \cdot \hat{\mathbf{n}}^2 & \mathbf{L}_2^2 \cdot \hat{\mathbf{n}}^2 & \dots & \mathbf{L}_n^2 \cdot \hat{\mathbf{n}}^2 \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{L}_1^K \cdot \hat{\mathbf{n}}^K & \mathbf{L}_2^K \cdot \hat{\mathbf{n}}^K & \dots & \mathbf{L}_n^K \cdot \hat{\mathbf{n}}^K \end{pmatrix} \tag{5.21}$$

and

$$\mathbf{M_{ij}} = \sum_k \left(\mathbf{L}_i^{(k)} \cdot \hat{\mathbf{n}}^{(k)}\right) \left(\mathbf{L}_j^{(k)} \cdot \hat{\mathbf{n}}^{(k)}\right) w^{(k)} \tag{5.22}$$

In this form, parameter estimation can sometimes be ill-posed, since the measurements made might not supply enough information to distinguish between different sets of transformations. Since by far the most common transformations in video sequences are translations, one can regularise by introducing a prior which is added to **M** to penalise non-translational modes (as in [251]). The equation is then solved using singular value decomposition. If **M** is singular (or close to being singular to the available precision), this method provides the closest $\boldsymbol{\alpha}$ that minimises $|\mathbf{M}\boldsymbol{\alpha} - \mathbf{v}|$ in the least squares sense.

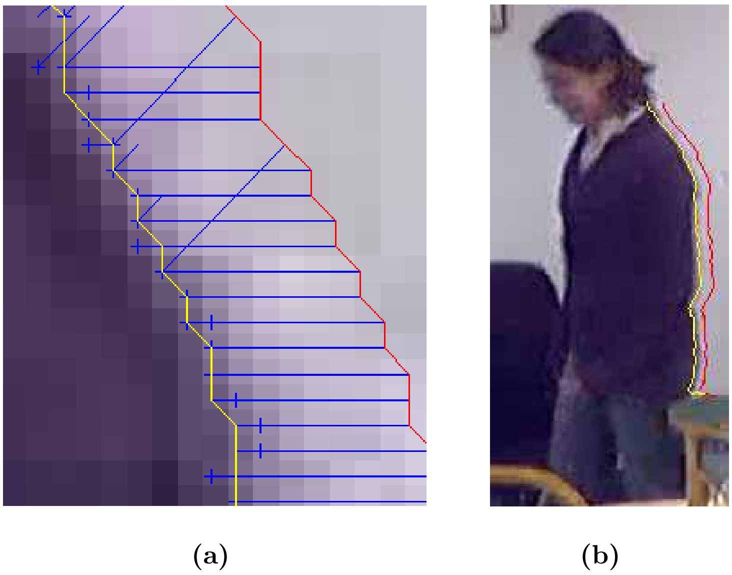


Figure 5.12: Example motion model. The red lines show the edge location in the previous frame and the yellow lines indicate the motion model's final positioning of the edge for the current frame. a) Close-up of the edgel matching with matching pairs illustrated using blue lines. b) The complete edge.

Figure 5.12 illustrates the process. Results can be significantly improved by also re-finding the edgels at each iteration of the motion model estimation. When measuring **d**, a search for the edgels is made from their current positions according to the motion model.

Factors such as large inter-frame motion, edge rotations by $\frac{\pi}{2}$ (which means that searching along the normal may produce few matches), or an unfortunate combination of motion and edge-shape, could mean that few pixel-matchings occur. The estimation might converge to an erroneous solution due to the fact that not enough matches have occurred to sufficiently restrict the possible solutions. However, by also repeating the edgel-refinding as the motion model is improved one can expect to get gradually improving matchings too, which will produce a better motion model. As suggested in [251], additional improvements are possible by integrating edge statistics over multiple frames.

5.6.3 Gradient Vector Flow Active Contours

Active Contours [23], also known as "snakes", are fitted to boundaries through minimisation of an energy function. A snake can be represented as a curve $\mathbf{c}(s) = [x(s), y(s)]$ with $s \in [0, 1]$ such that the snake's shape and position are adapted dynamically by minimising an energy term:

$$E_{\text{snake}} = E_{\text{internal}} + E_{\text{external}} \tag{5.23}$$

where

$$E_{\text{internal}} = \int_0^1 \frac{1}{2} \left[\alpha \left| \frac{d\mathbf{c}(s)}{ds} \right|^2 + \beta \left| \frac{d^2\mathbf{c}(s)}{ds^2} \right|^2 \right] \tag{5.24}$$

Internal forces E_{internal} penalise bending and promote smoothness and continuity through parameters α and β which control the snake's tension and rigidity respectively. The external energy E_{external} can be derived from image properties such as edges in order to force the snake towards salient image features or object boundaries. It can be written as the negative gradient of a potential function, i.e. $-\left|\bigtriangledown I(x,y)\right|^2$. However, active contours initialised with this external force have problems if initialised too far from an edge since the force drops sharply with increased distance from the edge. It is possible to increase the capture range by smoothing with a Gaussian kernel, however this process has a detrimental effect on edge localisation and strength.

Xu and Prince [297] propose a new external force field, the *Gradient Vector Flow* (GVF) field $\mathbf{v}(x, y) = [u(x, y), v(x, y)]$, which is the vector field that minimises the energy functional

$$\mathcal{E} = \int\int \mu(u_x^2 + u_y^2 + v_x^2 + v_y^2) + |\bigtriangledown \mathbf{f}|^2 \, |\mathbf{v} - \bigtriangledown \mathbf{f}|^2 \, dxdy \tag{5.25}$$

where u_x represents the partial derivative of u with respect to x, and $f(x, y)$ is an edge-map derived from the image. In areas close to an edge, i.e. when $|\bigtriangledown f|$ is large, the second term dominates the integrand. This means that in the vicinity of an edge, the energy $\mathcal{E}$ is minimised by setting $\mathbf{v} = |\bigtriangledown \mathbf{f}|$ as desired. In regions far from an edge, where $|\bigtriangledown \mathbf{f}|$ is small, the sum of the squares of the partial derivatives dominates the summation, hence providing a slowly-varying field which will still push the snake towards the edge.

However, the rate and quality of convergence is still improved by a better initialisation provided by the blob-tracker. The blob-tracker derives a foreground image where all pixels except those belonging to detected blobs are set to zero, and the edge-detector can then be used to build an edge map for this image. This is ideal for calculating the vector field, since there are very few edges except for those of the moving objects. The snake is then initialised around the convex hull of each blob and iteratively moved under this vector field such that it converges around the blob's outline, thus providing a smoothed and parameterised model of the complete outline of the object.

5.6.4 Integration

The blob-tracker provides an initial model of the position, appearance, and motion of objects. Edge detection and region segmentation allow one to correlate and refine the shape of detected blobs. Using the active contour approach, object shape is then parameterised as a closed boundary contour. By adapting the theoretical framework of the sample-based edge tracker, one can ascribe a parametric motion model to the object boundaries and internal object structure by tracking sample points along the contour and internal edges. The resulting framework models appearance, motion, and position of each tracked object in a way which is suitable for further stages of analysis.

Region segmentation for blob refinement

Blob detection relies on adaptive thresholding to detect and group pixel outliers from the background process. In order to improve the correspondence between detected blobs and actual objects (or object parts), the overlap between detected blobs and segmented regions is computed. Regions that largely overlap with candidate blobs are added to the corresponding blob. This process therefore produces blobs that are made up of regions bounded by edges.

Snake convergence to moving object boundaries

The next stage in the combined approach is to initialise the snake around the actual blob perimeter. Rather than only using edges to calculate the vector field, completed region boundaries are used as generated by the static segmentation, which improves snake convergence in areas where edges are broken. This results in a refined model of the object's outline, parameterised by an active contour.

Sample-based motion estimation of object boundaries and edges

The blob-tracker provides a prediction (calculated using Kalman or particle filtering) of the velocity of the centre of gravity of the object, which can be used to find an expected inter-frame translation for the whole outline. This is used to initialise the motion model for the edge-tracker, which has three main effects. Firstly, one can expect a better matching, and hence faster convergence of the motion model to occur. Secondly, one can apply the method to sequences with large inter-frame motions without having to increase the search distance in the edge-tracker. Thirdly, one generally no longer requires the translation prior since the bulk of the translation element has already been included in the model.

5.6.5 Shape Model

While the steps described above yield a robust parameterisation of the motion and contour of tracked objects, a more concise representation of the object's shape is required to ground the ontology as described in section 5.7.

A large number of shape models have been proposed in the literature [161, 30], which can be broadly classed into region and boundary based approaches. In the present case, region-based shape attributes such as centre of gravity, area, bounding box ratio, and principal axis orientation are already derived through the blob tracking framework. The main benefit of the edge and contour based approach described above lies in accurately localising the object's boundary, and hence the shape model considered here provides a concise parameterisation of that boundary.

The approach chosen here uses the first four of the seven affine invariant moments proposed by Hu [119]. For a continuous function f over two dimensions x and y, the moment m_{pq} of order $p+q$ $(p, q \in \mathbb{N})$ is defined as:

$$m_{pq} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^p y^q f(x,y) dx dy \tag{5.26}$$

which in the discrete case becomes

$$m_{pq} = \sum_x \sum_y x^p y^q f(x,y) \tag{5.27}$$

For a digital image, x and y are pixel coordinates and f is a function such as the intensity at a given pixel. The centroid of a shape can then be approximated by computing

$$\bar{x} = \frac{m_{10}}{m_{00}}; \ \bar{y} = \frac{m_{01}}{m_{00}} \tag{5.28}$$

over its pixels. To provide invariance with respect to translation, scale, and rotation, one can compute the centralised moments:

$$\mu_{pq} = \sum_x \sum_y (x - \bar{x})^p (y - \bar{y})^q f(x,y) \tag{5.29}$$

and then normalise as follows:

$$\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}^{\lambda}} \tag{5.30}$$

where $\lambda = \frac{(p+q)}{2} + 1$ and $(p+q) \geq 2$. From this Hu [119] derived a set of seven invariant moments ϕ_i, the first four of which are:

$$\begin{aligned}
\phi_1 &= \eta_{20} + \eta_{02} \\
\phi_2 &= (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2 \\
\phi_3 &= (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2 \\
\phi_4 &= (\eta_{30} - \eta_{12})^2 + (\eta_{21} - \eta_{03})^2
\end{aligned}$$

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.

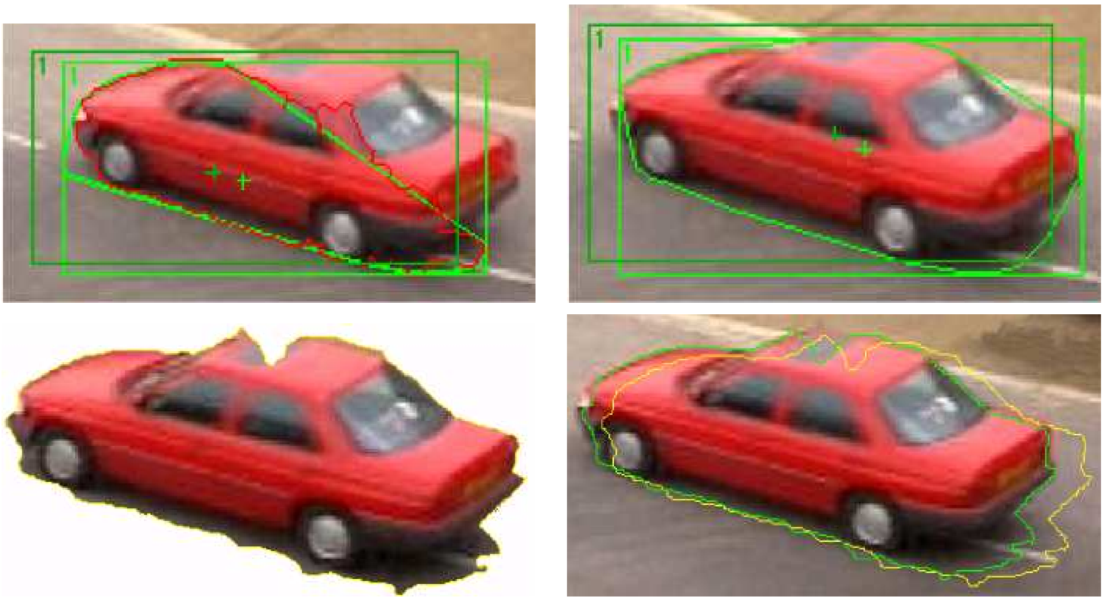


Figure 5.13: Walk-through of the combined approach. The blob corresponding to the tracked object (top left) is improved by using the static segmentation (top right) and a snake is initialised around the object (bottom left). In the second frame, the edge-tracker builds a motion model for the snake (bottom right).

Invariant moments provide a useful and concise parameterisation of shape [7] and can be computed efficiently. A number of refinements such as Zernike moments have been proposed in the literature [161, 30]. As before, the choice in the present context has been motivated by conciseness, generality, simplicity and clarity rather than optimality for any particular visual analysis application. Only the first four Hu moments are used since these proved adequate for the present purposes. More advanced shape measures, such as the recently proposed Cyclic Hidden Markov models introduced by Magee [165], could be used instead and may offer advantages for particular problem domains.

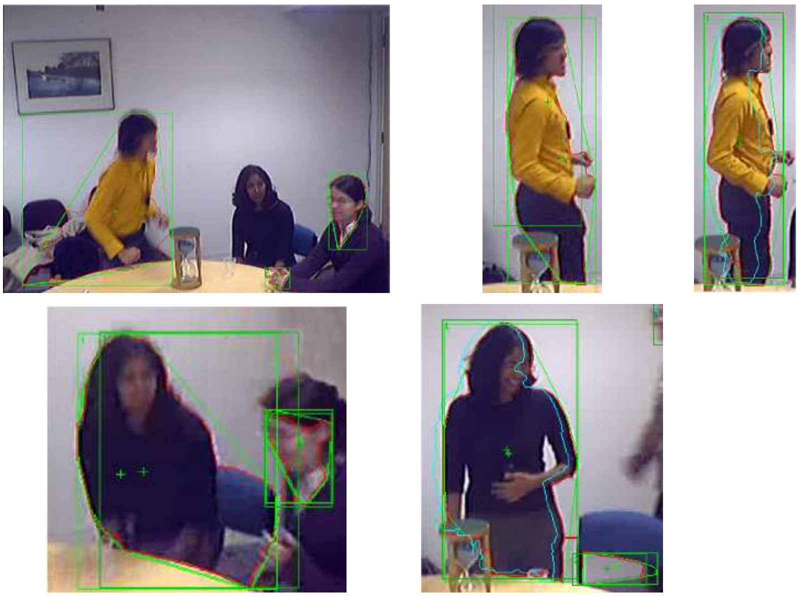


Figure 5.14: Example of combined blob, edge and contour based tracking for indoor surveillance. As the man and woman stand up, they are tracked by the blob detection and their contour is parameterised by an active contour. The edge-tracker then determines a motion model for the objects.

5.6.6 Tracking Results

Walk-through

The various stages of the process are demonstrated using a publicly available sequence from the *International Workshop on Performance Evaluation in Tracking and Surveillance* (PETS2000).

In figure 5.13, the green rectangle represents the bounding box of the blob (the dark green rectangle shows the predicted bounding box for the next frame),

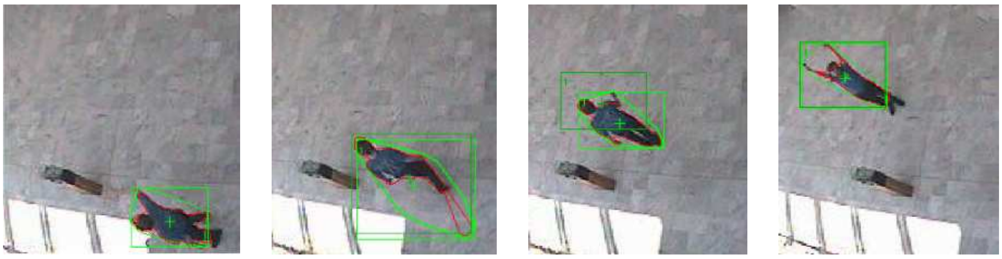


Figure 5.15: Results for the CAVIAR sequences. The light green rectangle indicates the object's bounding box while the dark green rectangle indicates the predicted object location for the next processing stage.

whilst the green polygon represents its convex hull. A GVF snake (shown in red) initialised on the convex hull of the blob fails to capture the true outline of the object. By contrast, initialisation of the snake around the segmentation-enhanced blob demonstrates a good bound on the car. Finally, the edge-tracker tracks the snake and produces a motion model.

Indoor surveillance

People are considerably more difficult to track and parameterise than rigid objects such as cars. Figure 5.14 shows results for a meeting-room scenario. As the man in the yellow shirt rises from the table, he leaves behind a "ghost object"[4] around which the snake is also initialised. However, as the blob-tracking framework identifies and removes the ghost, the improved outline of the man is located. A motion model is successfully calculated by the edge-tracker as the man sways backwards slightly. The woman's moving face is detected well. However, her hair is not grouped as part of a moving object since it does not change significantly in terms of colour. These frames show how given a decent initialisation, the snake can provide an excellent bound on the object.

CAVIAR Sequence

Figure 5.15 shows results for the tracking of a person in one of the CAVIAR sequences when sampling over every 5th frame only. The system is able to maintain

[4] Areas of background that are uncovered by a moving object may result in erroneous foreground detections, so called "ghosts", especially if the object had previously remained static sufficiently long to be incorporated into the background model.

a stable lock on the object despite large changes in edge shape.

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
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 5.3: Extended set of computer vision derived nodes

5.7 Extended Ontology

Based on the more sophisticated visual analysis methods presented in section 5.6, the CAVIAR ontology shown in table 5.1 was extended with a new set of visual descriptors. As can be seen from table 5.3, this includes the same descriptors introduced in section 5.5.1 plus additional ones for object shape, motion, deformations, and position. As suggested by the results below, position information can be a powerful cue for recognising some types of behaviour such as fighting, since these tend to occur in the same general region of the scene in the CAVIAR sequences.

The variables in table 5.3 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

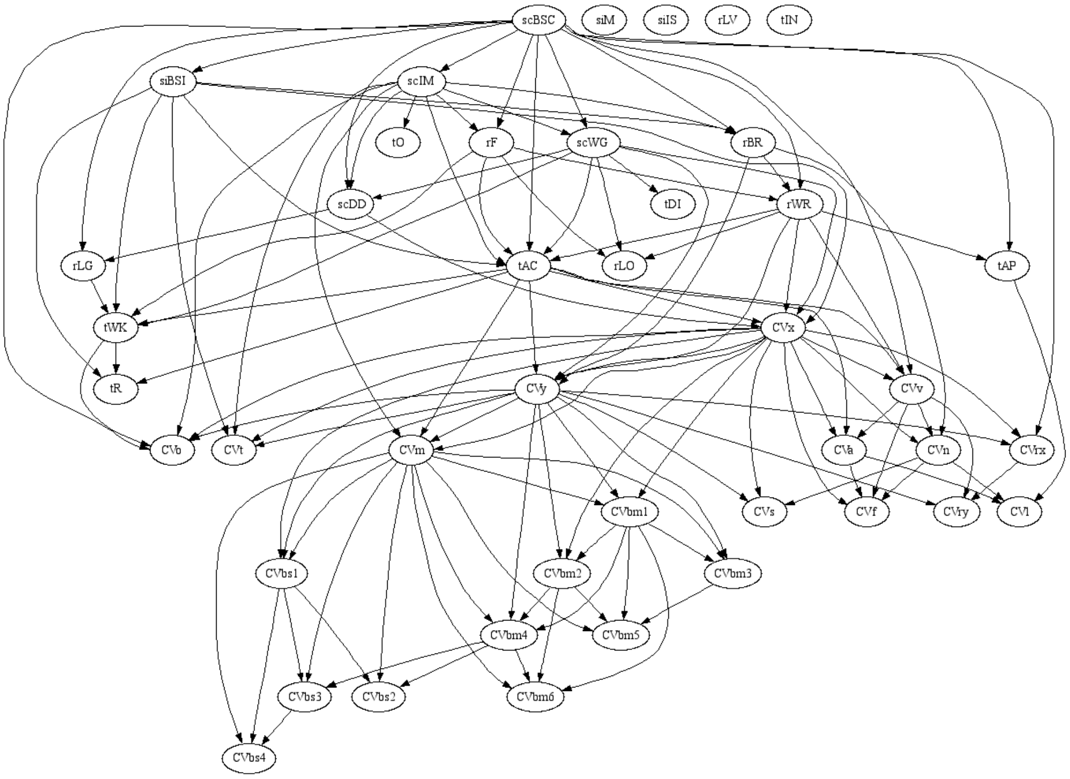


Figure 5.16: Bayesian network structure trained using the K2 algorithm from the extended ontology.

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

Figure 5.16 shows a Bayesian network trained for the new data set in the manner described in section 5.5.2 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

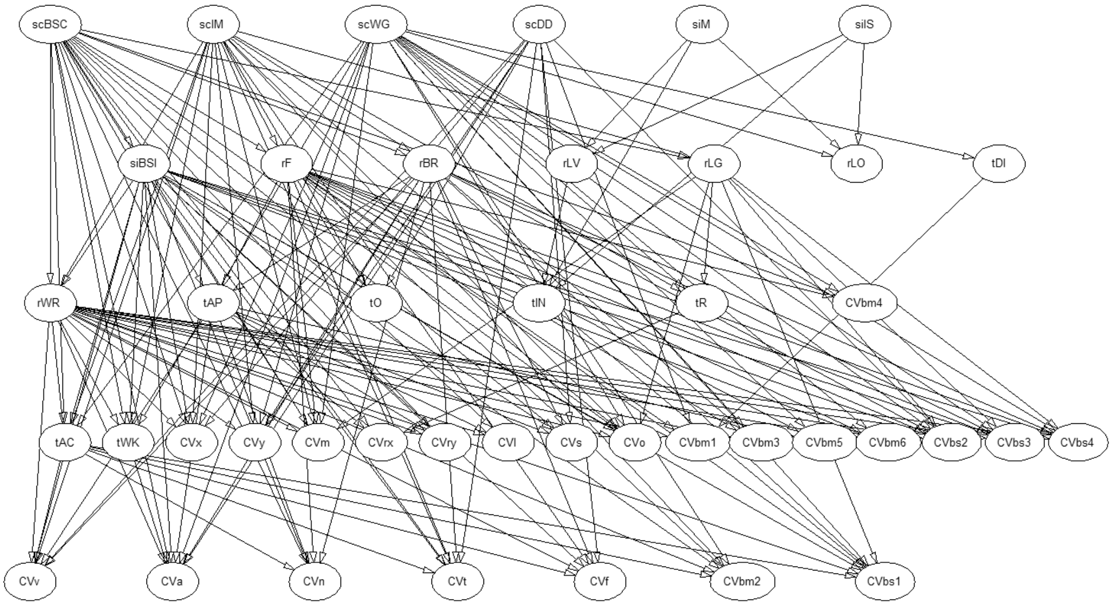


Figure 5.17: Bayesian network structure trained using the K2 algorithm with a structural prior from the extended ontology.

because their values are almost constant in the data set and hence can in most cases be inferred trivially through a purely deterministic prior.

The network shown in figure 5.17 was trained using K2 with a structural prior as described before. Figure 5.18 shows classification rates achieved by this network given different sets of evidence. As shown in figure 5.19, this network achieves better recognition rates than the one trained without a structural prior. However, the differences are now less pronounced and the achieved performance much higher than those reported in section 5.5.3 using the less sophisticated computer vision grounding methods. The average recognition rates 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 5.20 shows an example of a Bayesian network that was trained using K2 without any such ontological information, i.e. with a random initial ordering of the nodes. The

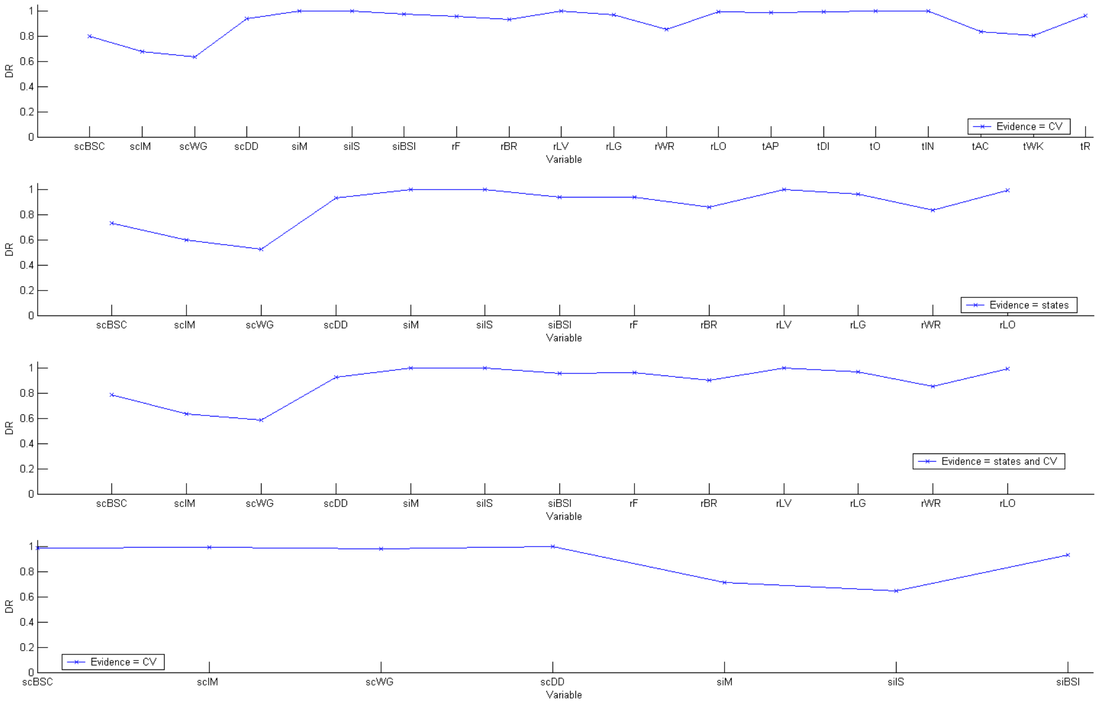


Figure 5.18: Plot of recognition rates achieved by the Bayesian network shown in figure 5.17 for the variables in the CAVIAR ontology given different evidence. ***Top*****: given only the computer vision derived nodes as evidence;** ***2nd from top*****: given only the states (tAP, tDI, tO, tIN, tAC, 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.**

performance of this network is worse than the two discussed before, with the average recognition rate now being 0.865.

Figure 5.21 shows a Bayesian network trained on the extended ontology using the MCMC algorithm as described in section 5.5.2. Although about 15000 MCMC sampling iterations were required to achieve acceptable convergence (following an initial "run-in" period of 5000 cycles), the MCMC method now also achieves better results than were previously obtained with the simpler set of visual descriptors. It is also possible to provide a structural prior by initialising the search performed by MCMC with a particular network structure. Figure 5.22 shows a network trained using a MCMC process whose starting point was the network structure shown in figure 5.17. Though superficially similar, the directed acyclic graphs of two

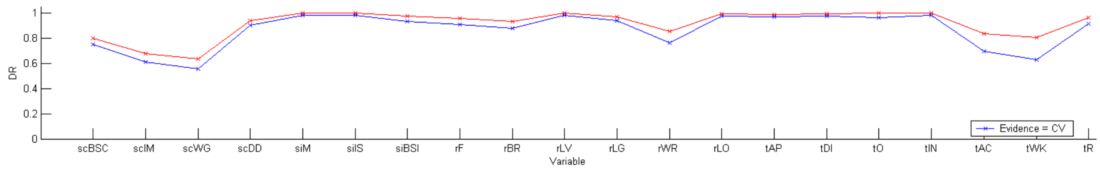


Figure 5.19: Plot of Bayesian network recognition rates for the variables in the CAVIAR ontology given only the computer vision derived nodes as evidence. The rates achieved by the network in figure 5.17 are shown in red, those for the network in figure 5.16 are shown in blue.

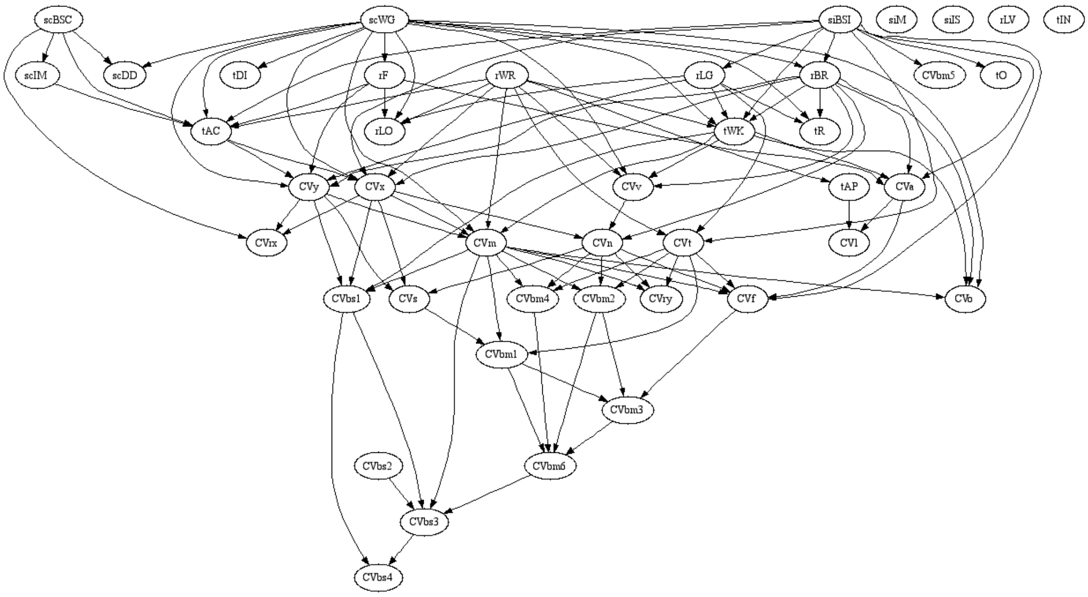


Figure 5.20: Bayesian network structure trained using the K2 algorithm with random initial ordering of the nodes.

network structures obtained by MCMC in this way differ in 226 entries of their connectivity matrices. The resulting average recognition rates for the CAVIAR ontology are 0.642 (no initialising prior, figure 5.21) and 0.659 (search initialised with a prior, figure 5.22).

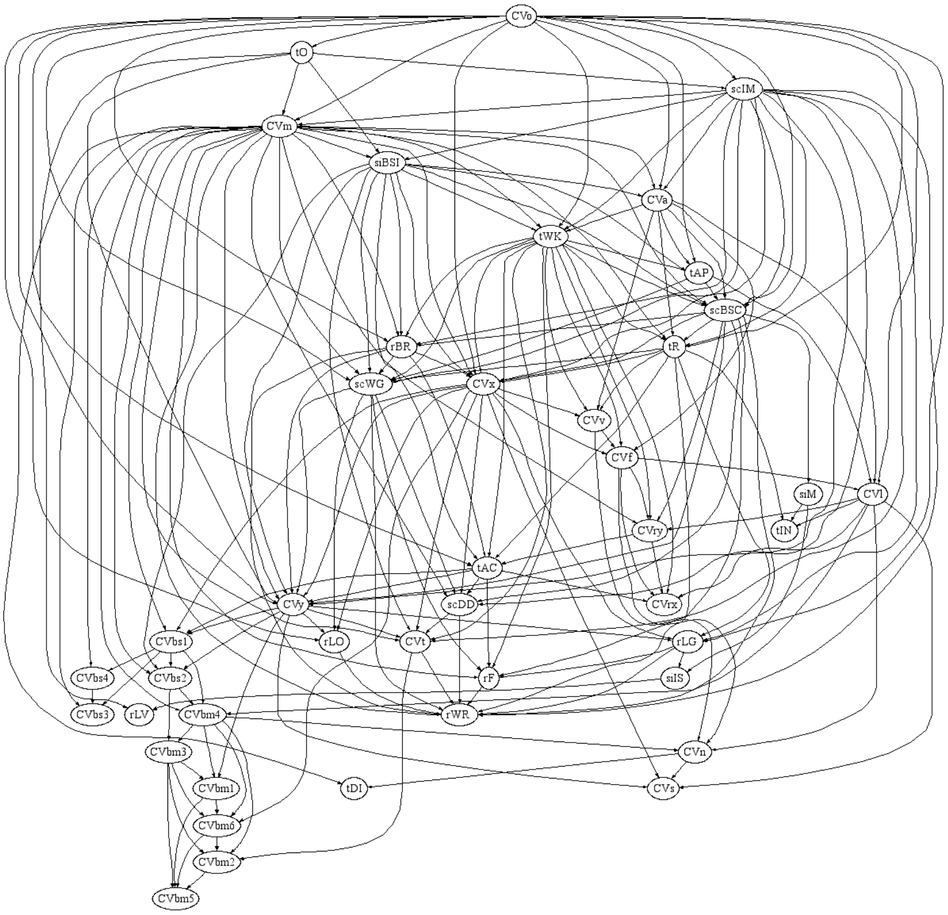


Figure 5.21: Bayesian network structure trained using the MCMC algorithm from the extended ontology.

5.8 Summary

This chapter demonstrates the value of using ontologies to build working high-level vision systems. It shows how such ontologies can be derived from an existing ground truth schema and a set of visual tracking methods. Together with a set of annotations, such an ontology can then be used to derive training data and prior structural constraints for automated learning of both the connectivity and parameters of Bayesian networks for high-level inference. Provided sufficient amounts of such data are available, this process is reasonably robust to human errors in the annotations and inadequacies in the automatically extracted visual content

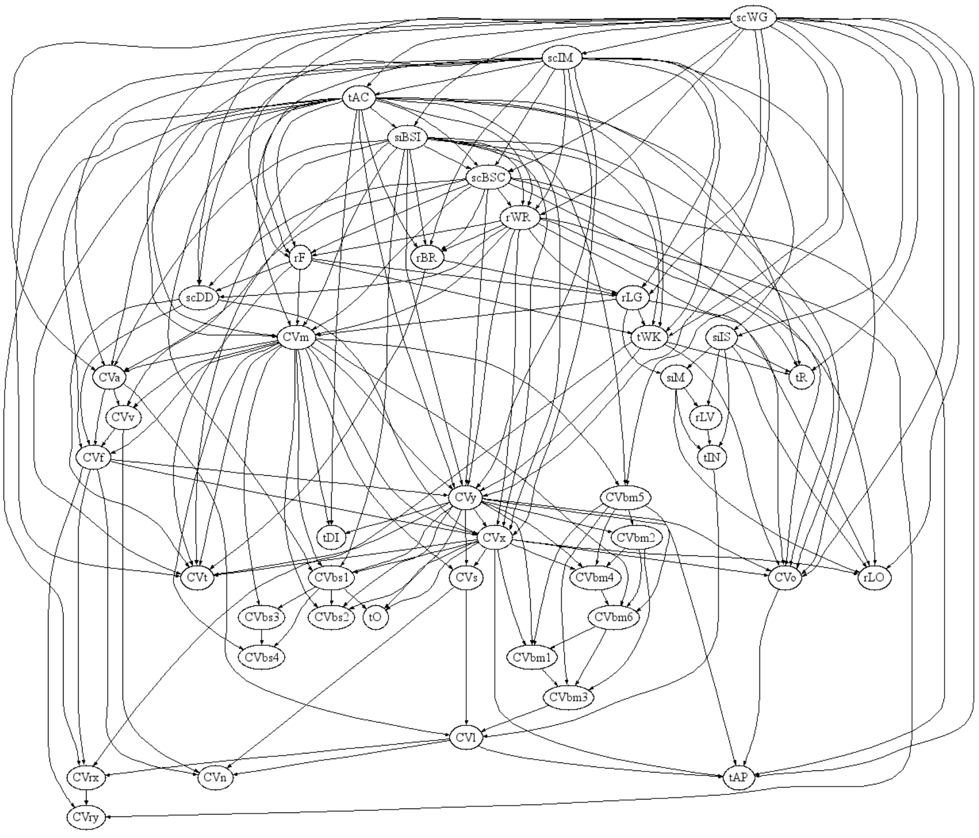


Figure 5.22: Bayesian network structure trained using the MCMC algorithm initialised with a structural prior from the extended ontology.

descriptions.

Performance analysis of the resulting networks and the quality of the visual tracking provides a useful basis for comparison of alternative schemes and methods. It allows alternative ontologies to be compared for their self-consistency and realisability in terms of the different visual detection and tracking modules.

It was shown how this can be done using the CAVIAR project's ground truth annotated sequences. Using the annotations and a robust but straightforward blob-based tracking framework, one can gather co-occurrence statistics for the terms in the ontology and use these together with the hierarchical relationships implied by it (event, scenario, situation, role, state) to build Bayesian networks for high-level analysis tasks in the chosen visual surveillance domain.

Chapter 6

Multi-sensory and Multi-modal Fusion for Sentient Computing

6.1 Overview

This chapter considers a multi-sensory and multi-modal fusion problem in which computer vision information obtained from calibrated cameras is integrated with a sentient computing system known as “SPIRIT”. The SPIRIT system employs an ultrasonic location infrastructure to track people and devices in an office building and model their state. It is shown how the resulting location and context data can be fused with a range of computer vision modules to augment the system’s perceptual and representational capabilities.

Vision techniques include background and object appearance modelling, face detection, segmentation, and tracking modules. Integration is achieved at the system level through the metaphor of shared perceptions, in the sense that the different modalities are guided by and provide updates to a shared world model. This model incorporates aspects of both the static (e.g. positions of office walls and doors) and the dynamic (e.g. location and appearance of devices and people) environment. The shared world model serves both as an ontology of prior information and as a language of context for applications. Fusion and inference are performed by Bayesian networks that model the probabilistic dependencies and reliabilities of different sources of information over time. This chapter shows that the fusion process significantly enhances the capabilities and robustness of both sensory modalities, thus enabling the system to maintain a richer and more accurate world model.

Section 6.2 discusses the concept of sentient computing, gives references to related work, and presents an outline of the SPIRIT system. In section 6.3 the problem of integrating visual information with the SPIRIT system is introduced and the computer vision methods employed are presented. Modelling of visual appearance of tracked entities is described in detail in section 6.4. The fusion of several vision algorithms with information from the SPIRIT system is then presented in section 6.5 and algorithms for combined multi-object tracking are described. Section 6.6 shows how these techniques can be applied to augment the representation of dynamic aspects of the SPIRIT world model while section 6.7 presents results for enhanced modelling and visualisation of the office environment.

6.2 Sentient Computing

6.2.1 Problem Definition and Context

Efforts in ubiquitous computing [290] are increasingly focused on providing a model of computing in which the proliferation of relatively cheap communications, sensors, and processing devices is leveraged in such a way as to make the resulting systems aware of aspects of their environment and the interactions which take place within it. The goal of what is termed *sentient*[1] [117] or *context-aware* [237] computing is to enable systems to perceive the world and relate to it in much the same way as people do, thereby creating the illusion of a shared perception that carries with it an implicit understanding of *context*. Indeed it can be argued that deriving an accurate representation of context is a “holy grail” of human computer interaction [65], as it would allow people to interact much more naturally with computer systems in a way which is pervasive and largely transparent to the user.

Sentient computing thus aims to model aspects of the context within which human-computer interactions take place in order to better infer and anticipate user intentions and requirements. This is achieved by integrating information from a range of networked sensors and processors distributed throughout a (typically

[1] From *sentient:* having the ability to perceive via the senses.

indoor) space in order to maintain an internal representation, or *world model*, of that environment. Applications utilise the world model in order to obtain implicit knowledge of user context. To realise the goal of shared perception, the robustness and accuracy of sensory data and its interpretation must approximate that of human beings in the chosen domain [76], and the world model must maintain an accurate up-to-date representation of context.

As section 6.2.2 will briefly describe, there are a number of sensor technologies and modalities that have been employed in building sentient computing systems. Since vision is our primary sensory modality, it too has attracted interest from researchers wishing to build sentient computing systems. However, each technology exhibits its own drawbacks, which intrinsically limit the capabilities of the overall system. Much recent work has therefore focused on combining different sensory modalities, often to the effect that one sensor system is found to be most reliable with additional complementary modalities augmenting its performance.

The sentient computing system considered in this chapter (see section 6.2.3) uses ultrasound to track tagged devices such as computers and phones within an office. It is currently the most accurate large-scale wireless tracking system of its kind. Nevertheless, systems of this kind have a number of limitations arising from the fact that they are largely restricted to tracking the 3D location of ultrasonic tags which must be attached to objects of interest. On the other hand, visual information holds the promise of delivering richer representations of the world without an inherent need to tag salient entities. Computer vision offers a range of capabilities such as detection, classification, and tracking, which are important prerequisites of a context-aware computing system [55]. However, apart from the need to deploy sufficient numbers of cameras to ensure adequate coverage, machine vision remains hampered by problems of *generality* and *robustness*, which reduce its suitability as a primary (or sole) sensory modality. There is clearly much scope for work that integrates information from these disparate sources.

6.2.2 Related Work

Efforts in ubiquitous and context-aware computing have employed a range of sensor systems [246, 115] such as accelerometers, touch sensitive surfaces, and more commonly sound (microphones, ultrasound) or light (infrared, radio) at different frequencies in order to detect or track people and devices.

Most indoor location systems such as the infrared-based Active Badge system [289] only allow positioning at room accuracy. Exceptions include the RADAR [11] project, which uses a series of base stations transmitting in the ISM radio band to provide an indoor location system with a median resolution in the range of 2 to 3 meters. The Cricket system [212] uses a series of ultrasonic beacons placed throughout a building at known locations, which periodically emit both ultrasonic pulses and a radio signal. Devices can use the time difference of the two pulses to determine the closest beacon and estimate their distance from it to within 5cm to 25cm. Much current research focuses on the use of Ultra-Wideband (UWB) radio positioning [154] and other wireless technologies such as Bluetooth and 802.11b. The requirements for low-cost sensor systems in applications such as retail stock control and supply chain management have fostered interest in cheap passive tagging technologies such as Radio Frequency IDentification (RFID) tags [288].

Outdoor location is a more challenging problem due to the lack of a controlled environment and suitable infrastructure. Most systems derive positioning information from the Global Positioning System (GPS) [90], which permits spatial resolutions of about 10m, although greater accuracy is achievable through the use of pre-calibrated reference points. The use of information from differential timing and directional signal strength variation with respect to multiple base stations allows handset localisation to within a few hundred metres of the widely used GSM mobile phone system [69]. Given the intrinsic drawbacks of any given positioning technology, sensor fusion approaches are gaining popularity [116]. A number of architectures have been proposed for this purpose, e.g. [190].

The role of computer vision in practical office-scale sentient systems has largely been restricted to the detection of tagged objects [127, 126], although vision-based systems for gesture recognition and motion analysis have been under development for some time [295, 88]. Approaches relying on probabilistic integration of different sources of visual evidence such as face detectors and models for motion, shape, and colour have shown particular promise (e.g. [217, 242, 256, 296]). However, the difficulties of deploying perceptual user interface mechanisms on the basis of vision alone are exacerbated by problems such as brittleness and lack of real-time performance and generality. Many vision-based systems have consequently been developed for fairly circumscribed control tasks where a limited vocabulary of pointing and selection gestures is appropriate. A truly pervasive system however requires a richer basis for interactions and a notion of context strong enough to recognise when no intended interaction with the system occurs.

These problems have led researchers to investigate fusion of vision with other sensory modalities. Most such systems rely on integration with sound in the audible range via microphone assemblies, which has proven particularly popular for videoconferencing applications [55, 301, 261, 56]. The EasyLiving project [31] integrates various sensors deployed in a single office. It primarily relies on a stereo camera visual tracking system that degrades badly as the number of people tracked causes errors due to occlusions and similarity in visual appearance.

A frequently cited research prototype called “KidsRoom” [24] provides a playroom that allows children to follow an interactive multimedia narrative. The system uses several cameras and a microphone to detect children and recognise certain classes of activity, but does not perform tracking. There are also systems that have integrated vision as a secondary modality to an existing system, for example [40], which deploys visual gait recognition as an additional identifying cue to a system based on a pressure sensitive “Active Floor” [5]. Multi-modal user localisation is also an important topic in robotics research [27], where information from stereo or omni-directional cameras mounted on a moving robot is often combined with sensors such as laser range finders.

Bayesian approaches to multi-modal fusion have been gaining prominence in the vision and other research communities. The approach presented in this chapter is related to work by Sherrah and Gong [242], which 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 using indicator variables to model observation reliabilities. Tracking on the basis of multiple sources of information is also demonstrated by Choudhury et al. [47], who present a system that fuses auditory and visual cues for speaker detection. Recent work by Beal et al. [17] demonstrates generative Bayesian fusion of microphone and camera data, while Torralba et al. [267] present work that highlights the importance of context for object and place recognition in situated vision systems.

6.2.3 The SPIRIT System

A sentient computing environment uses sensor and resource status data to maintain a model of the world that is shared between users and applications. Sensors and telemetry are used to keep the model accurate and up to date, while applications see the world via the model. A richer and more accurate model enables applications to better perceive context and thereby interact with users in a more natural way. For example, a call routing application could use location information to forward a call to whichever phone is closest to the intended recipient, but if that person appears to be in a meeting it may instead notify using a vibrating pager or forward the call to their voice mail facility or the office receptionist.

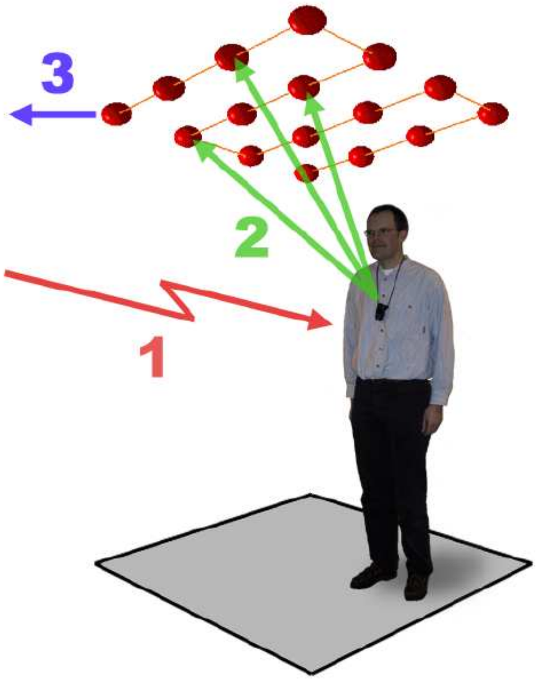


Figure 6.1: Function of the SPIRIT location system. A Bat sensor is triggered by radio (1), emits an ultrasonic pulse (2), and time-of-flight of the pulse is measured by receivers in the ceiling (3) to compute 3D position.

The SPIRIT[2] [6, 105, 4] project is a concrete implementation of these objectives. It was originally developed at AT&T Laboratories Cambridge, where it was used by 50 staff members and was in continuous operation for several years. The system is currently deployed throughout the Laboratory of Communication Engineering at Cambridge University (www-lce.eng.cam.ac.uk).

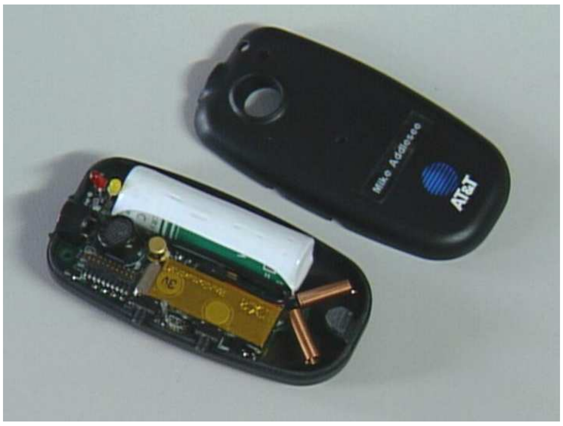


Figure 6.2: One of the wireless tracking tags known as a "Bat". The device is about the size of a matchbox and features two buttons for input and two LEDs and a sound chip for output. It contains a radio receiver, ultrasonic transmitter, microchip with a unique 48bit ID, and a AA lithium battery supplying enough power for up to 12 months of operation under normal conditions.

As shown in figure 6.1, the system uses mobile ultrasonic sensor devices known as "Bats" (shown in figure 6.2) and a receiver infrastructure to gather high-resolution location information for tagged objects such as people and machines. Such information is used to maintain a sophisticated world model of the office environment where it has been deployed. Applications can register with the system to receive notifications of relevant events to provide them with an awareness of the spatial context of user interactions. The achieved spatial granularity is better than 3cm for over 95% of Bat observations (assuming only small motion), and Bats may be polled using radio base stations with a variable quality of service to give update frequencies of up to 25Hz (shared among all Bats assigned to a given radio base station) while remaining scalable to hundreds of tagged people and devices

[2] Originally an acronym for "SPatially Indexed Resource Identification and Tracking".

in a large office. The Bats are equipped with two buttons, two LEDs and a sound chip to allow them to be used as portable input-output devices.

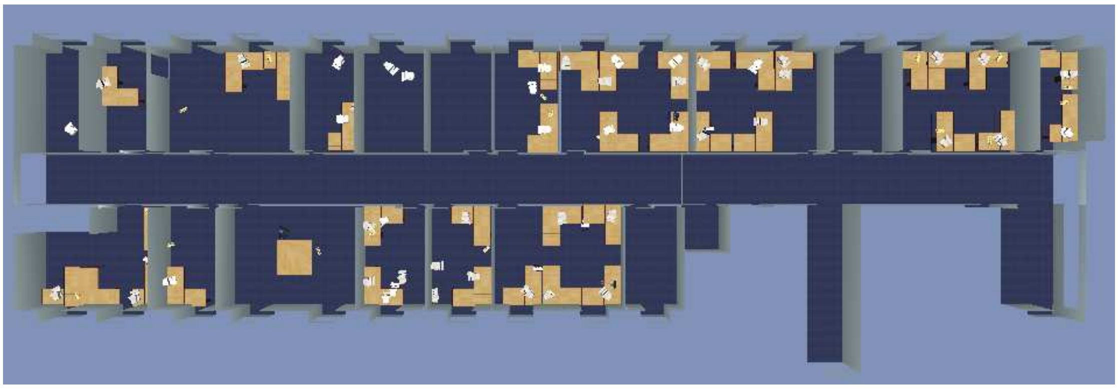


Figure 6.3: The "LabView" application displays a 3D real-time map representing the state of the world model. The bird's eye view shown provides an overview of the sentient office (at the Laboratory of Communications Engineering) and objects such as furniture, devices and people.

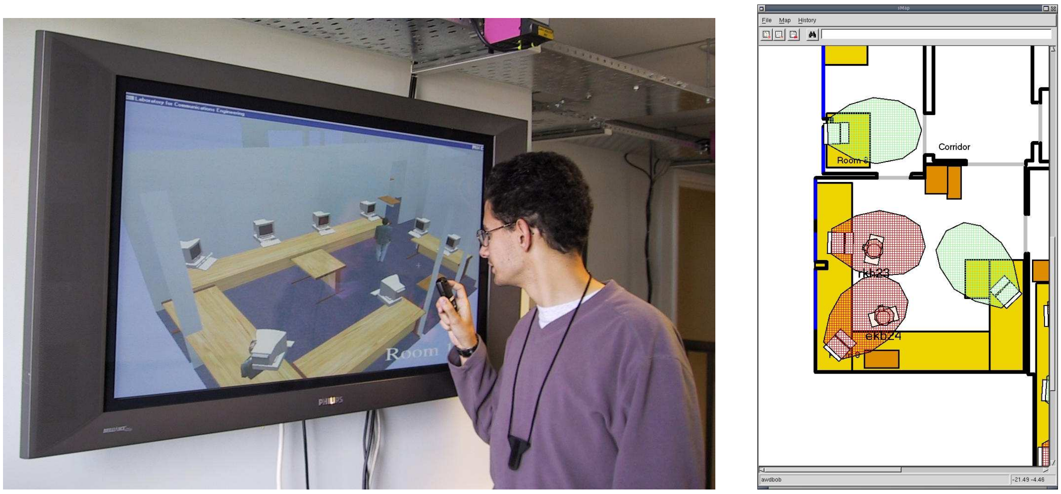


Figure 6.4: Interfaces for browsing the world model. ***Left:*** **Users can browse the LabView map by using their Bat as a 3D interface device.** ***Right:*** **The map is also available in a 2D view which can be used to edit parts of the world model and set spatial containment regions.**

The driving paradigm is that of "computing with space", i.e. physical location and spatial context (typically expressed in terms of containment and proximity) together with the attributes and capabilities of entities and devices present at a given time drive the behaviour of applications built upon the system. Some applications such as “LabView” shown in figures 6.3 and 6.4 allow users to navigate and browse the world model itself, while others respond to particular configurations of interest. Co-location and spatial composition can be used to infer aspects of context (e.g. “user A has entered office O”,“user B is using his Bat as a 3D mouse to control the scanner in corridor C”, “user B has picked up PDA P”) which can influence or trigger application behaviour, hence space itself becomes part of the user interface [117]. Current SPIRIT applications include “follow me event notification, personnel and resource localisation, office visualisation, user authentication, desktop teleporting, virtual 3D interfaces, and location support for augmented reality.

6.3 Integration of Visual Information

6.3.1 Motivation

Although the SPIRIT system has proven effective in providing fairly fine-grained spatial context upon which sentient computing applications can be built, difficulties remain. Bat system spatial observations are limited to the location of the Bat sensor, which is polled sporadically by a central base station. Each Bat has an associated identity (e.g. “Digital camera 1”, “User J.Smith”), which may carry associated semantics (e.g. digital cameras must be operated by a person, people can exhibit certain patterns of movement). However, only objects tagged with Bats can be tracked, and the model of the environment is static unless other sensors (e.g. light switches and temperature dials) provide information on it.

Computer vision methods can provide multi-modal human-computer interfaces with transparent detection, recognition, and tracking capabilities, but on their own suffer from a lack of robustness and autonomy in real world interaction scenarios. The integration of distinct sources of information about the world in light of application specific constraints holds great promise for building systems that can optimally leverage different sensory capabilities and failure characteristics.

Vision offers the possibility of acquiring much richer representations of entities in terms of their orientation, posture, and movements. It can also detect and to some extent classify and track additional features of the static (e.g. furniture) and dynamic (e.g. people and portable devices not equipped with Bats) environment. It may also be used to smooth over some of the difficulties inherent in an ultrasonic location infrastructure, thereby making it more robust. Information from the SPIRIT world model can in turn be used to provide constraints to the fusion process, to (re)initialise computer vision modules, and to act as a focus of attention mechanism.

6.3.2 Spatial and Temporal Correspondence

In order to fuse data from the visual and SPIRIT modalities, one must translate between their underlying representations. This requires translation between the 3D SPIRIT and 2D image reference frames and synchronisation of SPIRIT events with corresponding video frames acquired by a particular camera:

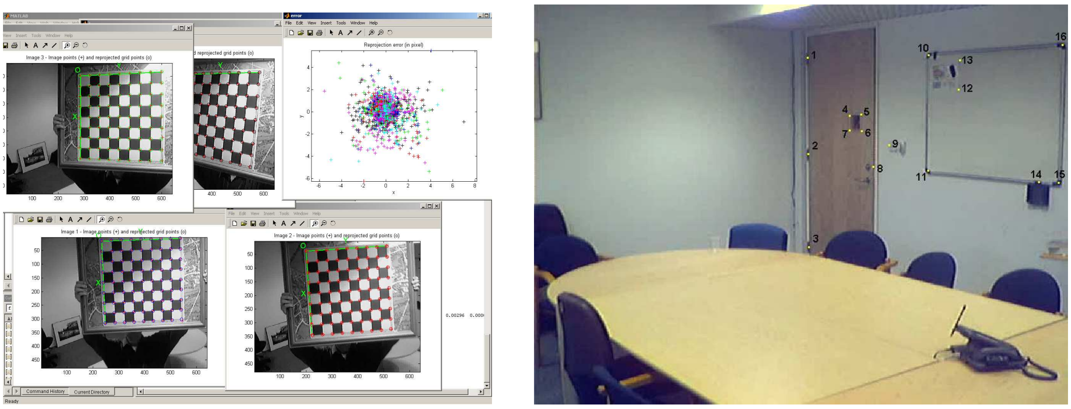


Image 3 - Image points (+) and reprojected grid points (o)
Reprojection error (in pixel)
Image 1 - Image points (+) and reprojected grid points (o)
Image 2 - Image points (+) and reprojected grid points (o)
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16

Figure 6.5: Calibration of intrinsic and extrinsic camera parameters. ***Left:*** **Use of the Matlab calibration toolbox for determining camera intrinsic parameters. Different views of a chessboard pattern are acquired by the camera and a numerical optimisation is performed to determine the parameters which can reproduce the projection to within a sub-pixel accuracy.** ***Right:*** **Several calibration points are marked and their positions determined in the 3D SPIRIT world frame and on the 2D image plane. A set of coplanar calibration points is sufficient to determine the projective mapping between the two coordinate systems.**

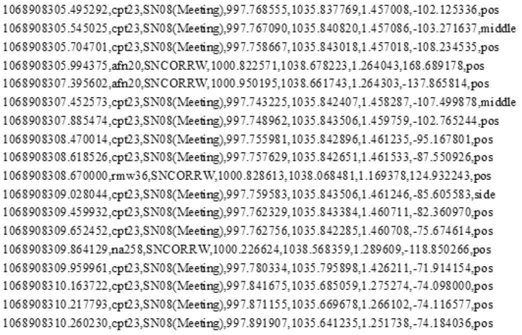
- *Frame of reference*: Visual input has been acquired from cameras placed at known locations within the SPIRIT world frame. Both the positions of the cameras and their intrinsic and extrinsic parameters were calibrated carefully. Intrinsic parameters were estimated using a chessboard calibration pattern and the Matlab toolbox developed by Jean-Yves Bouguet [29]. Several images (typically 20-30) of the chessboard were analysed, and camera parameters were estimated through correspondence analysis of positions of the corner points to establish planar homographies [298]. The camera model consists of 8 parameters [110] consisting of the coordinates of the 2x1 effective focal length and optic centre vectors, the skew coefficients accounting for non-orthogonality of the x-y axes, and four distortion coefficients representing radial and tangential distortions of the lens.

Camera position, view area, and extrinsic parameters were determined by means of a surveying application running on top of the SPIRIT system. This allowed feature points such as the position of the camera, the corners of its field of view, and calibration points visible by it to be localised very accurately in the 3D SPIRIT coordinate system. In order to determine the translation between 3D SPIRIT points and their 2D pixel coordinates when projected onto the camera's image plane, the image coordinates of the calibration points were also measured. After performing such measurements for a set of coplanar points[3], the extrinsic parameters consisting of a rotation matrix R and a translation vector T in 3D space can be determined numerically.

These steps make it possible to determine which objects should be visible (in the absence of occlusions) from a given camera, and to calculate the projection of 3D Bat system coordinates onto the image plane of that camera with a mean error of a few pixels. Figure 6.5 illustrates the calibration process.

- *Synchronisation*: SPIRIT location events need to be precisely synchronised with associated video frames. The synchronisation can be initialised manually using the buttons on the Bat device as described in figure 6.6. Arrival events for people entering the view area of a camera can be used to perform automatic re-synchronisation by using the visual tracking method and

[3] With the restriction that the plane on which they lie must not be parallel or orthogonal to the image plane to avoid unresolvable ambiguities.



```
1068908305.495292,cpt23,SN08(Meeting),997.768555,1035.837769,1.457008,-102.125336,pos
1068908305.545025,cpt23,SN08(Meeting),997.767090,1035.840820,1.457086,-103.271637,middle
1068908305.704701,cpt23,SN08(Meeting),997.758667,1035.843018,1.457018,-108.234535,pos
1068908305.994375,afn20,SNCORRW,1000.822571,1038.678223,1.264043,168.689178,pos
1068908307.395602,afn20,SNCORRW,1000.950195,1038.661743,1.264303,-137.865814,pos
1068908307.452573,cpt23,SN08(Meeting),997.743225,1035.842407,1.458287,-107.499878,middle
1068908307.885474,cpt23,SN08(Meeting),997.748962,1035.843506,1.459759,-102.765244,pos
1068908308.470014,cpt23,SN08(Meeting),997.755981,1035.842896,1.461235,-95.167801,pos
1068908308.618526,cpt23,SN08(Meeting),997.757629,1035.842651,1.461533,-87.550926,pos
1068908308.670000,rmw36,SNCORRW,1000.828613,1038.068481,1.169378,124.932243,pos
1068908309.028044,cpt23,SN08(Meeting),997.759583,1035.843506,1.461246,-85.605583,side
1068908309.459932,cpt23,SN08(Meeting),997.762329,1035.843384,1.460711,-82.360970,pos
1068908309.652452,cpt23,SN08(Meeting),997.762756,1035.842285,1.460708,-75.674614,pos
1068908309.864129,na258,SNCORRW,1000.226624,1038.568359,1.289609,-118.850266,pos
1068908309.959961,cpt23,SN08(Meeting),997.780334,1035.795898,1.426211,-71.914154,pos
1068908310.163722,cpt23,SN08(Meeting),997.841675,1035.685059,1.275274,-74.098000,pos
1068908310.217793,cpt23,SN08(Meeting),997.871155,1035.669678,1.266102,-74.116577,pos
1068908310.260230,cpt23,SN08(Meeting),997.891907,1035.641235,1.251738,-74.184036,pos
```

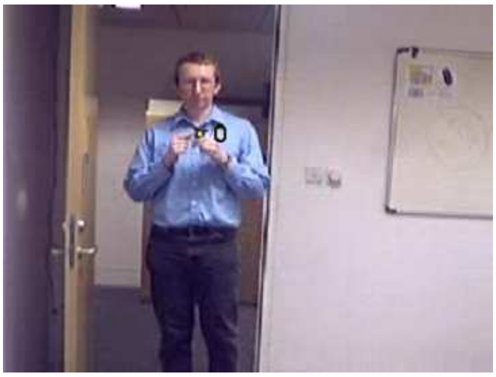


Figure 6.6: Synchronisation of SPIRIT event timestamps with corresponding video frames. *Left:* **Excerpt from a log file of location and "button press" events generated by the SPIRIT system.** *Right:* **By detecting a particular event such as a user pressing a button on his Bat device in front of a calibrated camera, one can associate the corresponding SPIRIT event timestamp with the video frame generated by the camera. The synchronous video stream and the asynchronous event stream can thus be synchronised at this point in time and subsequent events can be precisely ascribed to a particular video frame by noting their timestamps and converting this to a frame number given the frame rate of the camera.**

a motion history window to interpolate locations and correlate these to Bat sensor sightings.

Together with the camera calibration process described above, this enables data in both the spatial and temporal domain to be translated between the SPIRIT system and the visual information captured by the cameras. In order to fuse information from the two modalities, a number of additional issues are addressed in this chapter:

- *Quality of service*: Location information captured by a sensor infrastructure is a limited resource, and variable rates of quality of service are therefore imposed by the SPIRIT scheduler to determine the rate at which location events are generated for a given Bat. The frequency at which Bats are polled is reduced when the device is stationary. An internal accelerometer allows sampling rates to be increased when the Bat is in motion, or if it is to be used as a "3D mouse" to drive particular applications. However, there is some latency before an increased polling rate comes into effect.

- *Accuracy*: The accuracy of the SPIRIT location events is primarily affected by the properties of the sensor technology and the quality of the camera calibration and frame synchronisation. Ultrasound imposes intrinsic limits on update frequency and resolution due to its propagation characteristics and the filtering that is applied to dampen echoes and remove spurious (e.g. multi-path) observations.

- *Visibility*: The SPIRIT world model contains the locations of walls, doors, and windows, thereby making it possible to determine constraints on the environment viewed by each camera to deduce which objects and entities known to the system are likely to be visible by a given camera. A certain amount of occlusion reasoning may also be performed on this basis by computing a spatial ordering and predicting likely regions of overlap between tagged objects. However, there are many aspects of the world model such as furniture and the state of doors (open vs. closed) that are not fully modelled and must therefore be inferred by other means.

6.3.3 Vision Techniques for Detection, Identification, and Tracking

This section provides an overview of the vision techniques that have been implemented to provide additional information on the sentient computing environment and objects within it.

- *Skin colour classification*: Human skin colour is modelled as a region in HSV space defined by the following constraints taken from [87]:

  $$\begin{aligned} S &\leq 10;\ V \geq 40; \\ S &\leq -H - 0.1V - 110 \\ H &\leq -0.4V + 75 \\ \text{if } H \geq 0 \text{ then } S &\leq 0.08(100 - V)H + 0.5V \\ \text{else } S &\leq 0.5H + 35 \end{aligned}$$

  Histogram equalisation is applied to the entire image (or a target region predicted by other means) and candidate pixels that lie within the HSV subspace are clustered into regions using morphological operators to remove noise.

- *Face detection*: The face detection methods are applied to candidate regions identified by means of skin colour classification across the whole image or selectively to regions likely to contain faces as determined by the other modalities (i.e. head regions predicted by SPIRIT person observations or blob tracker based appearance models). In the former case, ellipse fitting is applied to the skin clusters and clusters may be split based on how elongated they are. Face detection is applied to PCA-transformed sub-windows of the candidate region at multiple scales.

  The face training set consists of about 1500 images from the CMU/MIT frontal face set rescaled to 25x31 and about 7000 non-face image patches randomly chosen from the Corel photo database. Two face detection methods were trained from this set: the first uses a generative mixture of Gaussians model trained using Expectation Maximisation and the second consists of polynomial kernel SVM classifiers. In both cases, the classifiers are arranged in a two-level cascade with the first classifier acting as a fast rejection filter for the second classifier, which was trained by incorporating test set misclassifications into the training set for the second stage. The two classification schemes are combined using simple disjunction of their binary classification

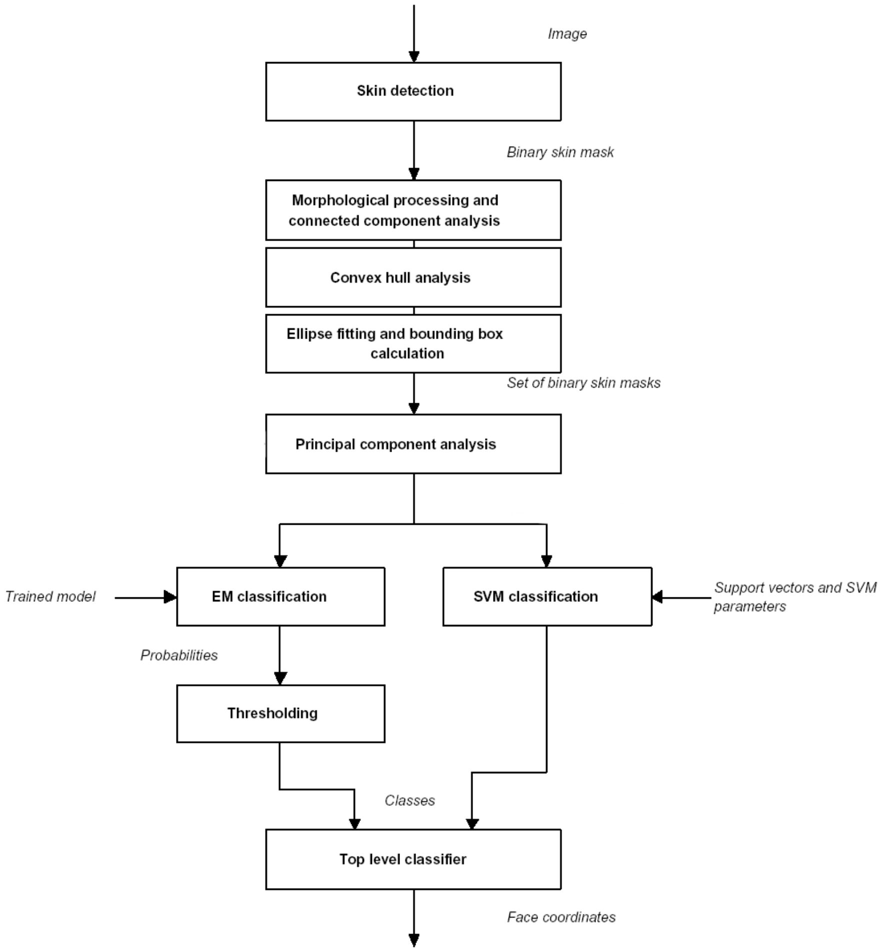


Figure 6.7: Overview of the face detection process.

decisions. This may increase false positive rates but ensures that fewer faces are missed. Figure 6.7 illustrates this method.

- *Background modelling and foreground detection*: As described in section 5.4.1, the system maintains a background model and foreground motion history which are adapted over time. The motion history M_t is used to identify a background image bim_t of pixels undergoing sufficiently slow change which can then be used to reliably update the background model B_t and estimate its variance. Pixels are deemed to be part of the dynamic foreground if they exceed a difference threshold that is a multiple of the background variance σ_t^B, and if they are not deemed to be part of a shadow as determined by

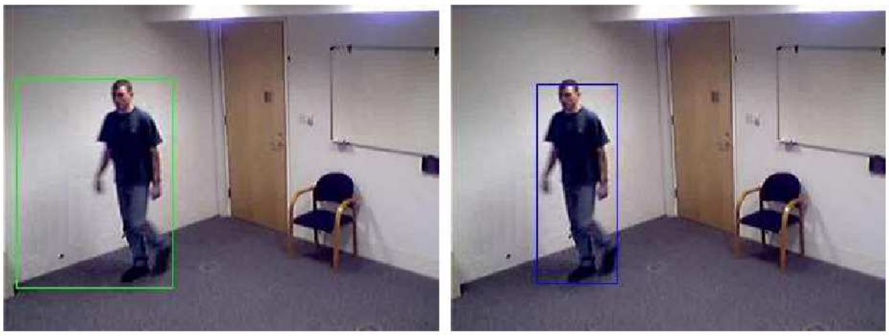


Figure 6.8: Sample result of the shadow detection method. In the image on the left, the person's shadow is incorporated into their blob as can be seen from the overly large bounding box. The image on the right shows the result after shadow detection has been applied to remove foreground pixels which are deemed to be part of a shadow. This results in a better blob and much tighter bounding box.

the DNM1 algorithm described in [211] (see section 5.4). Figure 6.8 gives an example of how shadow detection can improve blob detection.

Following these steps, candidate foreground pixels are subjected to morphological operations (dilation and erosion) to reduce noise in the final estimate.

- *Blob analysis and tracking*: Foreground pixels are clustered using connected component analysis to identify moving regions ("blobs"). These are then parameterised using shape (bounding box, centre of gravity, major axis orientation) and appearance measures as described in section 6.4. Blobs are tracked using a Kalman filter or Condensation tracker 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 shape and appearance. Figure 5.4 summarises the blob tracking framework.

- *Occlusion reasoning*[4]: To make tracking more robust, the object to blob assignment stage features a Bayesian network for reasoning about occlusions and object interactions (see figure 5.5) based on observed or predicted overlap of object bounding boxes and failures of object assignment. Dynamic

[4] See also section 5.4.3.

occlusions can also be disambiguated by using 3D SPIRIT data to predict spatial ordering of tracked objects, while static occlusions, object arrivals, and object departures can often be resolved with reference to the world model.

- *Static scene segmentation and region classification*: The region segmentation facilitates correspondence analysis between the world model and the scene viewed by a camera for environmental analysis and constrained tracking. As previously described in sections 4.5.2 and 4.5.3, the segmentation method due to [244] is applied to video frames at infrequent intervals. This method segments images into non-overlapping regions by computing a Canny-style colour edge detector and generating Voronoi seed points from the peaks in the distance transform of the edge image. Regions are 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. Sets of properties for size, colour, shape, and texture are computed for each region. These properties are fed into artificial neural network classifiers which have been trained to classify regions into "wood", "cloth", "carpet", and "internal walls". As explained in section 6.7.3, the classifiers were trained on images taken in the SPIRIT office and were found to perform well in identifying areas which would otherwise have been mislabelled (e.g. skin rather than wood) and in identifying furniture and wooden doors.

6.4 Vision-based Adaptive Appearance Modelling

6.4.1 Methods and Motivation

This section describes how fusion of three different appearance models enables robust tracking of multiple objects on the basis of colour information and by using the visual tracking framework described in section 6.3.3. In this chapter, short-term variation in object colour is modelled non-parametrically using adaptive binning histograms. Appearance changes at intermediate time scales are represented by semi-parametric (Gaussian mixture) models, while a parametric subspace method (Robust Principal Component Analysis, RPCA [61]) is employed to model long term stable appearance. Fusion of the three models is achieved through particle filtering and the Democratic integration method. It is shown how robust estimation and adaptation of the models both individually and in combination results in improved visual tracking accuracy.

Appearance models play a vital role in visual tracking of objects. They must remain robust to confounding factors such as noise, occlusions, lighting changes, and background variation, while adapting to appearance changes caused by motions and deformations of the tracked entity.

Global statistics such as colour histograms have been frequently used for object tracking due to their simplicity and versatility, see e.g. [191]. McKenna et al. [171] used Gaussian mixture models (GMM) to model the colour distribution of an object in order to perform tasks such as real-time tracking and segmentation. GMMs were shown to adapt over time to changes in appearance due to factors such as slowly-varying lighting conditions. Furthermore, many computer vision tasks can be posed as problems of learning low dimensional linear or multi-linear models. Principal Component Analysis (PCA) in particular is a popular view-based technique for parameterising shape, appearance, and motion [52].

Tracking algorithms which perform concurrent probabilistic integration of multiple complementary and redundant cues have been shown to be much more robust than those that utilise only a single cue [256, 242, 191, 268, 202]. In the particular case of colour and motion, the colour cue is the more persistent feature and can be used to maintain a lock on the object in the absence of motion. If the object is moving, the motion cue is the more discriminative feature and can be used (in preference to the colour cue) to keep track of the object.

Adaptive colour histograms can easily be completely re-estimated from frame to frame, and they are robust to the sort of short term noise and blur that would confuse the RPCA model. However, this may cause them to de-generate due to object motion or deformation. GMMs can be adapted selectively, and they combine aspects of both parametric and non-parametric estimation. Their explicit probabilistic interpretation via model likelihoods lends itself to incorporation in a wide variety of tracking and modelling frameworks. However they still suffer from some of the disadvantages of a global statistic. RPCA has stability due to its statistical outlier process, but is unable to cope well with short term changes in the object's appearance since these may appear as outliers. RPCA creates robust long term appearance models which can be used to re-acquire objects that have been temporarily lost due to occlusions or deformations.

Combining all three allows one to model intrinsic long term appearance (RPCA) as well as short term incidental changes (adaptive histogram, GMM) and expected variation of appearance due to object movement and gradual deformations. Much of the utility derives not from the models themselves but from the methods for matching and re-estimation (or adaptation). The important point about using appearance models for tracking is to model not only current appearance but also allowable (and hence expected) appearance variation.

6.4.2 Adaptive Binning Colour Histogram

Non-parametric density estimation techniques such as histograms assume no functional form for the underlying distribution and are robust to changes in orientation, relative position and occlusion of objects. Their simplicity and versatility make them suitable for modelling appearance over short time scales and during the initialisation phases of GMM and subspace estimation. The number of histogram bins is usually specified manually and remains fixed. If it is too small then the estimated density is very spiky, whereas if it is too large then some of the true structure in the density is smoothed out. In this chapter, the optimal value for the bin width is determined adaptively following the method of Leow et al. [157], who show that the mean error obtained by adaptive binning is about half that of fixed binning and that colour information is retained more accurately while requiring fewer bins.

The optimal number and width of histogram bins is determined using k-means clustering with colour differences computed in the CIELAB space using the CIE94 distance d_{kp} (see also [157]):

$$d_{kp} = \sqrt{\left(\frac{\Delta L^*}{S_L}\right)^2 + \left(\frac{\Delta C^*_{ab}}{S_C}\right)^2 + \left(\frac{\Delta H^*_{ab}}{S_H}\right)^2} \tag{6.1}$$

where ΔL^*, ΔC^*_{ab} and ΔH^*_{ab} are the differences in lightness, chromaticity, and hue between the centroid c_k of cluster k and pixel p with colour c_p. The values of the variables are $S_L = 1$, $S_C = 1 + 0.045\bar{\mathbf{C}}^*_{ab}$, and $S_H = 1 + 0.015\bar{\mathbf{C}}^*_{ab}$, where $\bar{\mathbf{C}}^*_{ab} = \sqrt{\mathbf{C}^*_{ab,k}\mathbf{C}^*_{ab,p}}$ is the geometric mean between the chromaticity values of c_k and c_p.

The clustering is repeated n times or until no pixels are left unclustered. To decrease computational cost, the number of blob pixels can be downsampled to some maximum number (usually 1000) prior to histogram adaptation.

Matching of tracked objects with candidate blobs is performed using weighted correlation. The similarity between two histogram bins is calculated by using a weighted product of the bin counts $H[i]$ and $H[j]$, where the weight w_{ij} is determined from the volume of intersection V_s between the two bins. Since the CIELAB space is perceptually uniform, histogram bins are spherical with radius r and so $V_s = V - \pi r^2 d + \frac{\pi}{12}d^3$ and $V = \frac{4}{3}\pi r^3$ where d is the distance between the bin centroids, therefore

$$w_{ij} = \frac{V_s}{V} = \begin{cases} 1 - \frac{3}{4}\frac{d}{r} + \frac{1}{16}(\frac{d}{r})^3 & \text{if } 0 \leq \frac{d}{r} \leq 2 \\ 0 & \text{otherwise} \end{cases} \quad ; \quad w_{ij} \in [0,1] \tag{6.2}$$

Dissimilarity of histograms H_p and H_q with n and n' bins respectively is then given by

$$D_{pq} = 1 - \sum_{i=1}^{n}\sum_{j=1}^{n'} w_{ij} H_p[i] H_q[j] \tag{6.3}$$

and normalised such that

$$\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} H_p[i] H_p[j] = \sum_{i=1}^{n'}\sum_{j=1}^{n'} w_{ij} H_q[i] H_q[j] = 1 \tag{6.4}$$

In order to incorporate some longer term appearance variation and smooth over fluctuations, the histograms are adapted using exponential averaging. Given a colour histogram H_t calculated for a blob at frame t and a smoothed object colour histogram S_{t-1} from frame $t-1$, the new smoothed object colour histogram S_t for frame t is given by $S_t = \alpha H_t + (1-\alpha)S_{t-1}$ where $\alpha = 1 - e^{-\frac{1}{\lambda}}$ determines the rate of adaptation. This is set to increase with increasing object speed in order to keep track of rapidly moving objects.

6.4.3 Gaussian Mixture Model

Gaussian mixture models (GMM) are a type of semi-parametric density estimation. Their use in colour modelling combines advantages of both parametric and non-parametric approaches. Most notably they are not restricted to certain functional forms (as for parametric approaches), and the model only grows with the complexity of the problem and not the size of the data set (as for non-parametric approaches). The conditional density for a pixel $\boldsymbol{\psi}$ belonging to an object O can be represented by a mixture of M Gaussians:

$$P(\boldsymbol{\psi}|O) = \sum_{j=1}^{M} P(\boldsymbol{\psi}|j)\pi(j); \quad \sum_{j=1}^{M} \pi(j) = 1; \quad 0 \leq \pi(j) \leq 1 \tag{6.5}$$

where the mixture parameters $\pi(j)$ give the prior probability that O was generated by the jth component and each mixture component is a Gaussian with mean $\boldsymbol{\mu}_j$ and covariance matrix $\boldsymbol{\Sigma}_j$:

$$P(\boldsymbol{\psi}|j) = \frac{1}{2\pi|\boldsymbol{\Sigma}_j|^{\frac{1}{2}}} e^{-\frac{1}{2}(\boldsymbol{\psi}-\boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1}(\boldsymbol{\psi}-\boldsymbol{\mu}_j)}$$

In this chapter, mixture modelling is performed in Hue-Saturation (HS) space to gain a degree of illumination invariance. Model estimation is performed on blob pixels using subsampling for efficiency and discarding samples whose intensity value is very low or close to saturation. Components are estimated using k-means with priors computed from the proportion of samples in each cluster. The parameters of the Gaussians (mean and covariance) are calculated from the clusters. Model order selection is performed using cross validation on a training and validation set randomly selected from the pixel samples. The training set is used to train a number of models of different order M, iteratively applying the Expectation Maximisation algorithm and splitting the component j with the lowest “responsibility” r_j for the validation set as given by

$$r_j = \sum_{\xi} P(j|\xi) = \sum_{\xi} \frac{P(\xi|j)\pi(j)}{\sum_{i=1}^{M} P(\xi|i)\pi(i)} \tag{6.6}$$

Component splitting involves creating two new components from an existing component, and then discarding the existing component. The process terminates once a maximum in the likelihood function is found or the maximum number of iterations has been exceeded.

Adaptation of the GMM over time is performed using the approach suggested in [171]. Given previous recursive estimates ($\boldsymbol{\mu}^{(t-1)}$, $\boldsymbol{\Sigma}^{(t-1)}$, $\pi^{(t-1)}$), the estimates derived for the new data ($\boldsymbol{\mu}^{(t)}$, $\boldsymbol{\Sigma}^{(t)}$, $\pi^{(t)}$), and estimates based on old data ($\boldsymbol{\mu}^{(t-L-1)}$, $\boldsymbol{\Sigma}^{(t-L-1)}$, $\pi^{(t-L-1)}$), the new mixture parameters for mixture model component j are derived thus:

$$\boldsymbol{\mu}^{(t)} = \boldsymbol{\mu}^{(t-1)} + \frac{r^{(t)}}{D^{(t)}}(\boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}^{(t-1)}) - \frac{r^{(t-L-1)}}{D^{(t)}}(\boldsymbol{\mu}^{(t-L-1)} - \boldsymbol{\mu}^{(t-1)}) \tag{6.7}$$

$$\boldsymbol{\Sigma}^{(t)} = \boldsymbol{\Sigma}^{(t-1)} + \frac{r^{(t)}}{D^{(t)}}(\boldsymbol{\Sigma}^{(t)} - \boldsymbol{\Sigma}^{(t-1)}) - \frac{r^{(t-L-1)}}{D^{(t)}}(\boldsymbol{\Sigma}^{(t-L-1)} - \boldsymbol{\Sigma}^{(t-1)}) \tag{6.8}$$

$$\pi^{(t)} = \pi^{(t-1)} + \frac{N^{(t)}}{\sum_{\tau=t-L}^{(t)} N^{(\tau)}}(\pi^{(t)} - \pi^{(t-1)}) - \frac{N^{(t-L-1)}}{\sum_{\tau=t-L}^{(t)} N^{(\tau)}}(\pi^{(t-L-1)} - \pi^{(t-1)}) \tag{6.9}$$

In the above equations, $N^{(t)}$ denotes the number of pixels in the data set, $D^{(t)} = \sum\limits_{\tau=t-L}^{(t)} r^{(\tau)}$, and $r^{(t)} = \sum\limits_{\xi \in X^{(t)}} P(j|\xi)$. The adaptivity of the model is controlled by the parameter L.

Matching of blobs to objects is performed by calculating the blob's normalised data log-likelihood $\mathcal{L}$ with respect to the object's GMM from a sample of blob pixels $X^{(t)}$ in the current frame:

$$\mathcal{L} = \frac{1}{N^{(t)}} \sum_{\xi \in X^{(t)}} \log P(\xi|O) \tag{6.10}$$

The log-likelihood threshold g for accepting a match is adapted over time to take into account current and previous log-likelihoods. Given an array of n most recent data log-likelihoods calculated for the previous n frames, the threshold is set to $g = \upsilon - k\sigma$, where υ is the median and σ is the standard deviation of the previous n data log-likelihood values.

6.4.4 Robust Principal Component Analysis

In order to acquire a stable model of object appearance over longer timescales, an extension of the Robust Principal Component Analysis (RPCA) method proposed by De la Torre and Black [61, 62] is applied. Given a matrix $\mathbf{D} = [\mathbf{d}_1 \mathbf{d}_2 \dots \mathbf{d}_n]$ whose column vectors $\mathbf{d}_i$ represent images each containing d pixels, the purpose of PCA is to find a lower dimensional subspace of k principal components $\mathbf{B} = [\mathbf{b}_1 \mathbf{b}_2 \dots \mathbf{b}_k]$ $(k \leq n)$ such that each image $\mathbf{d}_i$ can be approximated by

$$\mathbf{d}_i^{\sim} = \mathbf{B}\mathbf{B}^T \mathbf{d}_i = \mathbf{B}\mathbf{c}_i$$

where $\mathbf{c}_i$ are linear coefficients obtained by projecting the data onto the subspace:

$$\mathbf{C} = [\mathbf{c}_1 \mathbf{c}_2 \dots \mathbf{c}_n] = \mathbf{B}^T \mathbf{D}$$

PCA can be formulated as least-squares estimation of the basis images $\mathbf{B}$ by minimising

$$E_{pca}(\mathbf{B}) = \sum_{i=1}^{n} e_{pca}(\mathbf{e}_i) = \sum_{i=1}^{n} \left\| \mathbf{d}_i - \mathbf{B}\mathbf{B}^T \mathbf{d}_i \right\|_2 = \sum_{i=1}^{n} \sum_{p=1}^{d} \left(d_{pi} - \sum_{j=1}^{k} b_{pj} c_{ji} \right)^2 \tag{6.11}$$

where $c_{ji} = \sum_{t=1}^{d} b_{tj} d_{ti}$, $\mathbf{B}^T \mathbf{B} = \mathbf{I}$, and $e_{pca}(\mathbf{e}_i) = \mathbf{e}_i^T \mathbf{e}_i$ is the reconstruction error of $\mathbf{d}_i$.

RPCA enhances standard PCA by means of a pixel outlier process using M-estimators: Given n training images represented by matrix $\mathbf{D}$ as above and with scale parameters $\boldsymbol{\sigma} = [\sigma_1 \sigma_2 \dots \sigma_d]^T$, the error equation above is reformulated to obtain RPCA robust mean $\boldsymbol{\mu}$, bases $\mathbf{B}$, and coefficients $\mathbf{C}$:

$$\begin{aligned} E_{rpca}(\mathbf{B}, \mathbf{C}, \boldsymbol{\mu}, \boldsymbol{\sigma}) = & \sum_{i=1}^{n} e_{rpca} \left(\mathbf{d}_i - \boldsymbol{\mu} - \mathbf{B}\mathbf{c}_i \ , \ \boldsymbol{\sigma} \right) \\ = & \sum_{i=1}^{n} \sum_{p=1}^{d} \rho \left(d_{pi} - \boldsymbol{\mu}_p - \sum_{j=1}^{k} b_{pj} c_{ji} \ , \ \sigma_p \right) \end{aligned} \tag{6.12}$$

where ρ is the Geman-McClure error function $\rho(x, \ \sigma_p) = \frac{x^2}{x^2 + \sigma_p^2}$, and σ_p is a scale parameter that controls convexity and hence determines which residual errors are treated as outliers. To compute the mean and the subspace spanned by the first k principal components robustly, equation 6.12 is minimised using gradient descent with a local quadratic approximation. The parameters σ_p are updated during each iteration of gradient descent by estimating the median absolute deviation around each pixel (see [61] for details).

To ensure adequate performance for tracking, RPCA has been extended in this work using a robust incremental subspace learning technique to efficiently recompute the Eigenspace (see below). In addition, rather than computing RPCA over image intensity alone, two approaches were implemented to retain colour information:

- The simpler approach maintains separate RPCA subspaces for the hue, saturation, and luminance channels. Matching is performed through weighted summation of the Eigenspace Euclidean distances (see equation 6.19). Best results were achieved by weighting distance in hue space with 0.5, saturation with 0.3, and luminance with 0.2, again reflecting the desirability of discounting absolute brightness values to achieve illumination invariance.

- Secondly, RPCA was applied to one-dimensional colour statistics histograms derived from the colour distribution of each object in HSV space. Following Hanbury [102] a saturation-weighted hue histogram (where hue H is measured as an angle in the range $\{0°, 1°, \ldots, 360°\}$) is calculated by using the HSV saturation values as a weight differentiating between chromatic and achromatic (i.e. white, black or grey) colours

  $$W_\theta = \sum_x S_x \delta_{\theta H_x} \tag{6.13}$$

  where θ denotes a bin of the histogram over all pixel samples x with $\theta \in \{0°, 1°, \ldots, 360°\}$. S_x is the saturation of x, H_x the hue, and δ_{ij} is the Kronecker delta function. Alternatively, RPCA was also implemented for a saturation-weighted hue mean histogram $H_{S\ell}$ or saturation-weighted mean length histogram $R_{n\ell}$. Both of these histograms are calculated at each sample luminance level $\ell \in \{0, 1, 2, \ldots, N\}$:

$$H_{S\ell} = \arctan\left(\frac{B_{S\ell}}{A_{S\ell}}\right), \qquad R_{n\ell} = \frac{\sqrt{A_{S\ell}^2 + B_{S\ell}^2}}{\sum_x \delta_{L_x\ell}} \tag{6.14}$$

where H_x, S_x, and L_x are the hue, saturation, and luminance values at pixel location x and

$$A_{S\ell} = \sum_x S_x \cos H_x \delta_{L_x\ell}, \qquad B_{S\ell} = \sum_x S_x \sin H_x \delta_{L_x\ell}$$

RPCA based on the saturation-weighted hue mean histogram gave best results and is the method used in the experiments.

The number of pixels in the sample sets for Eigenspace computation was normalised by sub-sampling (and if necessary re-sampling) object pixels or through normalisation of the colour statistics histograms. Re-estimation of the RPCA model can be performed in batch mode by maintaining a moving window of previous samples (usually 10 or more). This approach was found to be cumbersome, and consequently a far more efficient incremental algorithm was devised by adapting the method proposed in [245] to re-estimate the RPCA coefficients. Incremental learning of the subspace parameters also has the advantage of increased robustness in the context of an online estimation problem such as that of appearance modelling for tracking. Given the current RPCA robust mean $\boldsymbol{\mu}^{(t)}$, bases $\mathbf{B}^{(t)}$, coefficients $\mathbf{C}^{(t)}$, and data sample $\mathbf{x}$, then at each frame t the algorithm proceeds as follows:

1. Project the data sample $\mathbf{x}$ into the current Eigenspace $\mathbf{B}^{(t)}$ and form the reconstruction $\mathbf{y}$ of the data:

$$\mathbf{c} = \mathbf{B}^{(t)\,T}(\mathbf{x} - \boldsymbol{\mu}^{(t)}); \quad \mathbf{y} = \mathbf{B}^{(t)}\mathbf{c} + \boldsymbol{\mu}^{(t)} \tag{6.15}$$

2. Compute the residual vector $\mathbf{r} = \mathbf{x} - \mathbf{y}$, which is orthogonal to $\mathbf{B}^{(t)}$, and form matrices $\mathbf{B}_e$ and $\mathbf{C}_e$:

$$\mathbf{B}_e = \left[\mathbf{B}^{(t)}\ \frac{\mathbf{r}}{||\mathbf{r}||}\right]; \quad \mathbf{C}_e = \begin{bmatrix} \mathbf{C}^{(t)} & \mathbf{c} \\ \mathbf{0} & ||\mathbf{r}|| \end{bmatrix} \tag{6.16}$$

3. Compute Robust PCA on $\mathbf{C}_e$, and obtain the updated robust mean $\boldsymbol{\mu}_s$ and robust bases $\mathbf{B}_s$. Discard the least significant Eigenvector of the new basis $\mathbf{B}_s$ and obtain the coefficient matrix for frame $t+1$:

$$\mathbf{C}^{(t+1)} = \mathbf{B}_s^T(\mathbf{C}_e - \boldsymbol{\mu}_s \mathbf{1}_{1\times(t+1)}) \tag{6.17}$$

where $\mathbf{1}_{m\times n}$ denotes a matrix of dimension $m \times n$ such that all the elements are 1.

4. Calculate the new basis matrix $\mathbf{B}^{(t+1)}$ and new mean $\boldsymbol{\mu}^{(t+1)}$ for frame $t+1$:

$$\mathbf{B}^{(t+1)} = \mathbf{B}_e\mathbf{B}_s; \quad \boldsymbol{\mu}^{(t+1)} = \boldsymbol{\mu}^{(t)} + \mathbf{B}_e\boldsymbol{\mu}_s \tag{6.18}$$

Results for the incremental algorithm suggest that it is at least an order of magnitude faster than the moving window technique for sufficiently large window sizes. It also offers the advantage of maintaining a more up-to-date Eigenspace since re-estimation occurs after every processed video frame.

In order to compute the match distance between a candidate blob represented by a column vector $\mathbf{e}$ of pixel samples and an object represented by RPCA parameters ($\mathbf{B}$, $\mathbf{C}$, $\boldsymbol{\mu}$), $\mathbf{e}$ is projected into the RPCA subspace. This is achieved by computing projection coefficients $\tilde{\mathbf{c}}$ which minimise

$$E(\tilde{\mathbf{c}}) = \sum_{j=1}^{d} \rho\left(e_j - \mu_j - \sum_{i=1}^{k} b_{ji}\tilde{c}_i \ , \ \ \sigma_j\right) \tag{6.19}$$

where ρ is the Geman-McClure error function as before. The object-blob distance is then defined as the minimum of the Euclidean distances between the blob coefficients $\tilde{\mathbf{c}}$ and each column of the object RPCA coefficient matrix $\mathbf{C}$.

6.4.5 Integration through Condensation

Particle filtering algorithms such as Condensation (conditional density propagation, [128]) pose the problem of tracking as estimation of states $\mathbf{X}$ from observations $\mathbf{Z}$ using the recursion:

$$P(\mathbf{X}_t|\mathbf{Z}_t) \propto \mathcal{L}(\mathbf{Z}_t|\mathbf{X}_t) \int P(\mathbf{X}_t|\mathbf{X}_{t-1})P(\mathbf{X}_{t-1}|\mathbf{Z}_{t-1})d\mathbf{X}_{t-1} \tag{6.20}$$

where the dynamical model $P(\mathbf{X}_t|\mathbf{X}_{t-1})$ describes state evolution and the observation likelihood model $\mathcal{L}(\mathbf{Z}_t|\mathbf{X}_t)$ gives the likelihood of any state in light of current observations.

The posterior probability distribution (6.20) is then represented by a weighted set of "particles"

$$P(\mathbf{X}_t|\mathbf{Z}_t) = \{s_t^{(n)}, \pi_t^{(n)} \mid n = 1 \ldots N\} \tag{6.21}$$

where $s_t^{(n)}$ is the nth sample and $\pi_t^{(n)}$ is the corresponding weight such that $\sum_n \pi^{(n)} = 1$. At each step of the Condensation algorithm the evolution of the weighted sample set is calculated by applying the dynamical model to the set. The observation likelihood function is then used to correct the prediction by calculating the weight π_t of each element in the set, i.e. $\pi_t \propto \mathcal{L}(\mathbf{Z}_t|\mathbf{X}_t^{(n)})$. N samples are then drawn with replacement by choosing a particular sample with probability $\pi^{(n)} = P(Z_t|X_t = s_t^{(n)})$. The mean state vector of an object in frame t is then modelled as the expectation

$$E[S] = \sum_{n=1}^{N} \pi^{(n)} s^{(n)}$$

Here, the observation density is modelled by a function that contains Gaussian peaks where the observation density is assumed to be high, that is, where an object could have generated a set of blobs with high probability. Each Gaussian peak corresponds to the position of a blob, and the peak is scaled by the object-blob distance. The likelihood $\mathcal{L}$ for a particle is computed as

$$\mathcal{L}(\mathbf{Z}_t|\mathbf{X}_t) \propto e^{-k \times \mathrm{dist}^2} \tag{6.22}$$

where dist is a distance under one of the appearance models of the local image patch at a given particle and the object under consideration, and k is a constant.

Likelihoods are calculated for each particle for each of the three appearance modelling schemes above and combined as follows:

$$\mathcal{L}(\mathbf{Z}_t|\mathbf{X}_t) \propto [\mathcal{L}_{rpca}(\mathbf{Z}_t|\mathbf{X_t})]^{\alpha_1} [\mathcal{L}_{chist}(\mathbf{Z}_t|\mathbf{X}_t)]^{\alpha_2} [\mathcal{L}_{gmm}(\mathbf{Z}_t|\mathbf{X}_t)]^{\alpha_3} \tag{6.23}$$

where $0 \leq \alpha_1, \alpha_2, \alpha_3 \leq 1$ are the reliability weights for each appearance model, initialised to $\frac{1}{3}$.

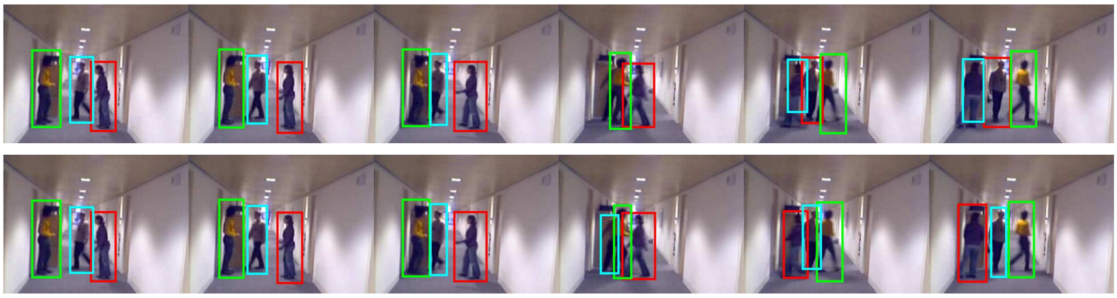


Figure 6.9: Indoor tracking results using vision cues only. The rectangles indicate bounding boxes of tracked objects and their colour indicates object identity assignment throughout the sequence. ***Top*****: tracking using only blob features and distances.** ***Bottom*****: tracking using the robust fusion of adaptive appearance models as described in section 6.4. Note how this allows identity of tracked entities (indicated by bounding box colour) to be maintained during and across occlusions.**

6.4.6 Adaptation of Cue Weights by Democratic Integration

Adaptation of the weights in equation 6.23 is performed dynamically during tracking by extending the idea of Democratic integration [274, 256] to the Condensation framework. Four separate observation likelihoods are computed: one for the joint appearance model, and three for each of the RPCA, adaptive histogram and GMM appearance cues. Condensation is performed separately for each of the observation functions, resulting in four hypotheses, R_{fused}, R_{rpca}, R_{chist}, and R_{gmm}, which are regions where the object is thought to be in the current frame. Each region centroid is obtained by computing the expectation of the respective particle sets for each cue.

The Euclidean distances $E_{k,t}$ between the centroid of R_{fused} and the centroids of R_{rpca}, R_{chist}, R_{gmm} at time t are then calculated. Since the joint observation function is assumed to exhibit the best performance, appearance cues that result in relatively large values of $E_{k,t}$ are considered less reliable in the current frame, and their reliability weight is lowered accordingly. A score $\gamma_{k,t}$ is computed for

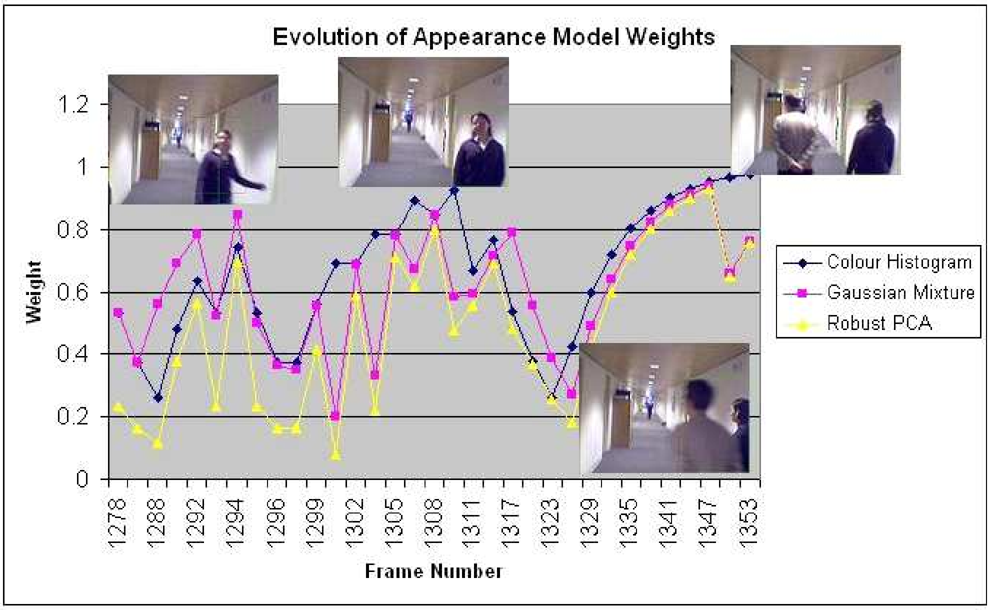


Figure 6.10: Evolution of reliability weights and object colour appearance models. ***Left*****: graph plotting the reliabilities of the appearance model cues for the woman shown in the test sequence. There is an initial rise in the reliability of all models due to the clear visibility of the woman. The large fall in reliability at frame 1320 onwards is due to occlusion by the man entering the scene. After the occlusion the appearance models successfully recover and their reliability increases very rapidly. Note the lag of the RPCA (and in some cases the Gaussian mixture) model behind the colour histogram model due to their slower adaptation.**

each cue k as follows:

$$\gamma_{k,t} = \frac{\tanh(-aE_{k,t} + b) + 1}{2} \tag{6.24}$$

where a, b are constants (set to 2 and 5 respectively) and tanh is the hyperbolic tangent function. Given $\gamma_{k,t}$, the weights $\alpha_{k,t}$ for each cue k are then adapted using first order exponential averaging:

$$\alpha_{k,t+1} = \beta\gamma_{k,t} + (1 - \beta)\alpha_{k,t} \tag{6.25}$$

where β controls the rate of adaptation (setting $\beta = 0.75$ was found to give good results in most sequences). Performing the Condensation algorithm four times during each frame was found not to be a bottleneck since most of the computation time is required for the particle distances (which need only be computed once per frame).

6.4.7 Results

To evaluate the adaptive appearance models and the fusion mechanism discussed above, testing was carried out on a number of indoor surveillance sequences which were acquired in the same office in which the SPIRIT system has been deployed (see section 6.5.2). The tracking conditions are especially demanding due to the presence of intermittent bright lighting, skin coloured walls, motion blur and occlusions as the people interact. Figure 6.9 shows how the fusion framework makes tracking robust with respect to occlusions and movements of people (the results shown are for sequence *S7* discussed in section 6.5.2). In figure 6.10 it is shown how the appearance modelling improves accuracy in light of erroneous blob hypotheses generated by the background differencing and blob detection framework.

Quantitative results for multi-hypothesis tracking using visual cues alone are presented in section 6.6.1 and contrasted with those achievable through integration of SPIRIT information as described in section 6.5.

6.5 Multi-modal and Multi-sensory Fusion

6.5.1 The World Model as a Shared Ontology

The SPIRIT system maintains an internal dynamic representation of the office environment including objects and events within it. This world model comprises a static part consisting of those aspects of the environment which are not monitored using the ultrasonic location system, and a dynamic part consisting of those objects and devices which are. The former includes the location and spatial extent of rooms, walls, windows, doors, and items of furniture such as desks and shelves that were manually added to the model. The latter tracks personnel, visitors, portable devices, and other assets that have been tagged with one or more of the Bats. Figure 6.11 gives an overview of the ontology of types and objects that is defined by the SPIRIT software system.

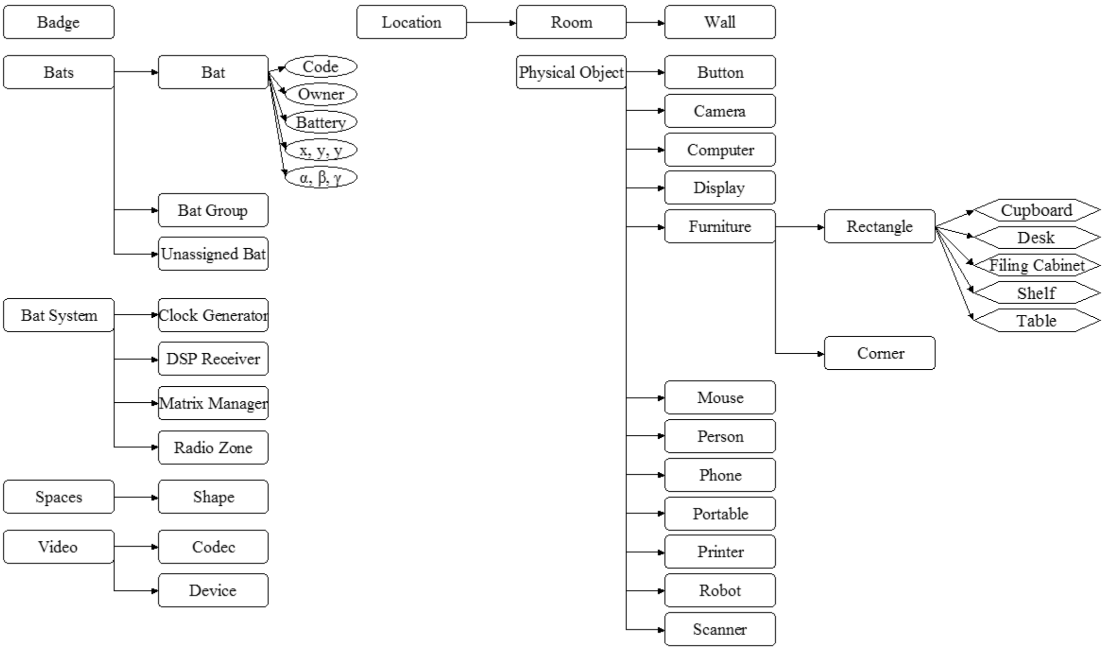


Figure 6.11: Simplified representation of the ontology of categories and entities defined by the SPIRIT system. For purposes of illustration, the list of key attributes associated with the "Bat" entity (unique code ID, owner, battery status, 3D spatial position and rotation) has been expanded and the subtypes of the "Rectangle" furniture subtype (cupboard, desk, filing cabinet, shelf, table) are shown.

The world model can thus be viewed as an ontology encompassing locations and spatial regions, objects (people, computers, phones, devices, cameras, furniture etc.), and event states (motions, spatial overlap, proximity, button events etc.). It serves as a language for context designed to keep the model consistent internally and with the sensor data. As shown in figure 6.12, applications see a description of the environment that is abstracted away from the sensor level.

The interpretation of such information is application dependent, for example routing a phone call may require finding the phone that is closest to a given person (provided they are in the same room), whereas a "follow-me" desktop session would need to know something about the user's orientation relative to available screens in order to select the best one for display.

Context thus determines the way in which activities are interpreted, and the perceived activities in turn may change the context. It is being continually defined

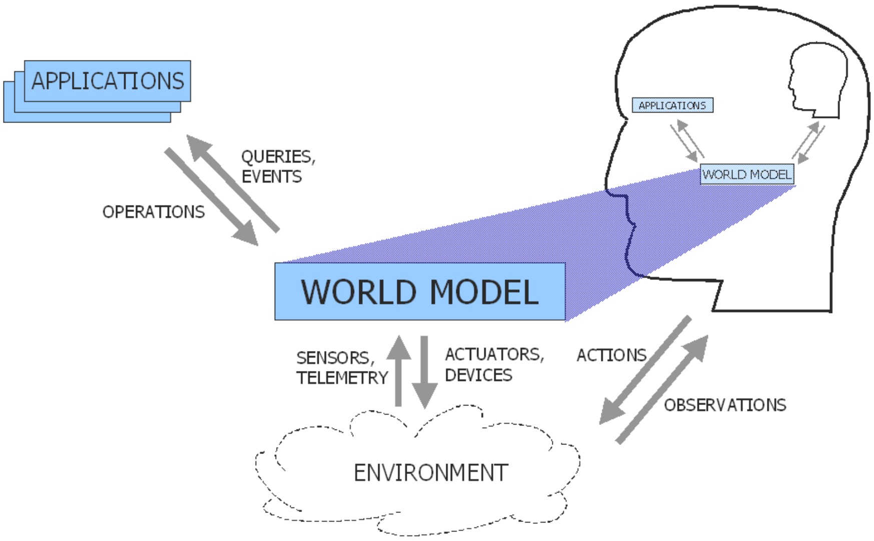


Figure 6.12: Diagrammatic overview of the world model maintained by the sentient computing system.

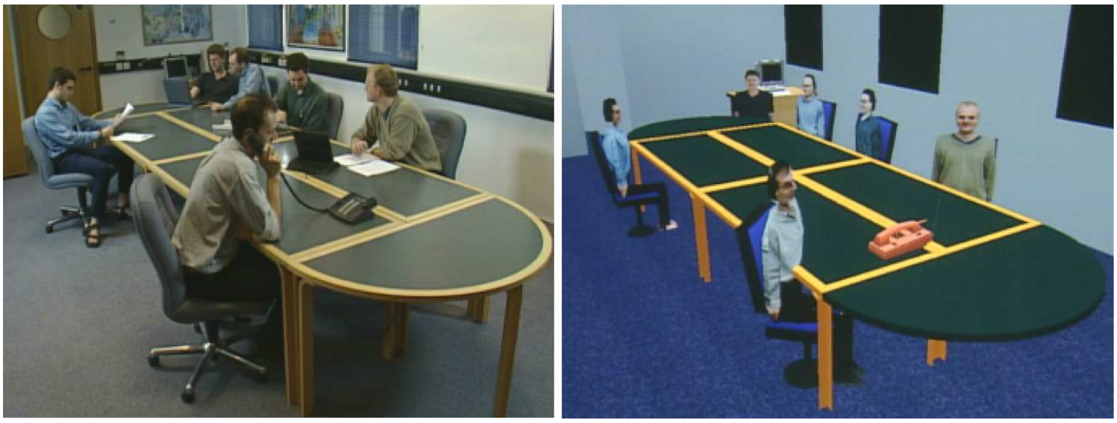


Figure 6.13: The world perceived by (left) users and (right) the sentient computing system (LabView visualisation).

and refined in the course of interaction and allows such interactions to become intelligible and meaningful. While the evolution of context and context-dependent interpretation is dynamic, a prior notion of what comprises models of context and means of inferring it is required. In the approach presented here, this prerequisite is expressed through an ontology consisting of the world model maintained by the SPIRIT system, which is augmented with information gained through a number of visual cues. Computer vision has for some time recognised the importance of integrating top-down and bottom-up information using knowledge hierarchies and domain constraints. Furthermore, the circular problem of defining context in terms of perceptions and interpreting these perceptions in terms of the given context naturally gives rise to a solution framework based on feedback.

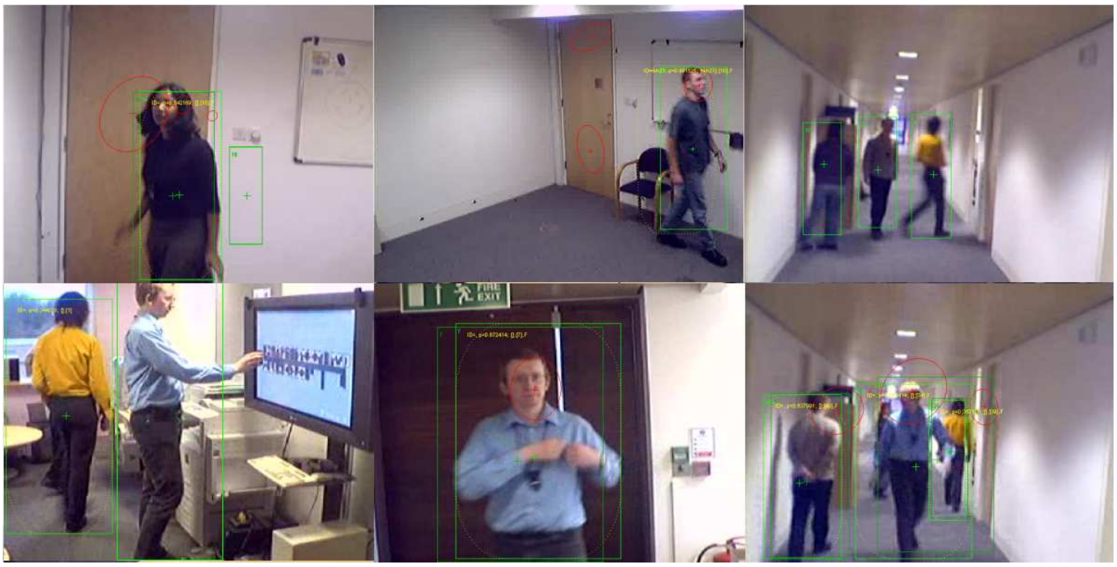


Figure 6.14: Example images from the test sequences (from left to right, top to bottom: sequence S9, S1, S7, S10, S8, S7). Green rectangles indicate bounding boxes of blob tracker objects, green polygons are the convex hulls of corresponding blob regions. Red ellipses are skin clusters, those with a red cross were deemed to contain a face. The cyan and magenta coloured rectangles are predicted object locations based on SPIRIT observations (magenta indicates an older observation). Yellow dotted ellipses and numbers indicate hypotheses resulting from the fusion process described below.

6.5.2 Experimental Setup

To facilitate the integration of visual information into the world model, cameras were deployed in various parts of the sentient office (see map in figure 6.3), namely one at the western entrance, one facing east along the corridor, and two in the meeting room (3rd room from the left at the bottom part of the map). Additional experimental sequences were also taken in some of the other rooms of the office.

The cameras used were two standard Philips webcams yielding a picture resolution and frame rate of (320x240 pixels, 12fps) and (640x480 pixels, 15fps) respectively.

A number of sequences featuring a total of 7 individuals were acquired to experiment with different scenarios and activities, such as one or more people entering or leaving a room, holding a meeting, drawing on the whiteboard, walking in front of each other, etc.. Several thousand frames from 10 sequences were manually annotated[5] by marking the bounding boxes of peoples' bodies and heads visible within each frame. A brief description of each follows:

- *S1*: One person enters the meeting room, walks around, and draws on the whiteboard. In addition to body outline and head region, the position of the person's hands was also annotated to allow for further research into gesture recognition and related applications.

- *S2*: Two people enter the meeting room one after the other, greet each other, and walk around the room at different speeds, frequently occluding one another.

- *S3*: Two people hold a meeting, entering and leaving the room one after the other. At various points they sit around the meeting room table, get up, draw on the whiteboard, and pass objects to one another.

- *S4*: The same scene depicted in sequence S3, but viewed from a different camera mounted on a different wall and at different height from that in S3.

- *S5*: A similar scenario as in S2 and S3, but involving up to five people simultaneously. The participants enter and eventually leave the room one after the other. They perform actions such as greeting one another, swapping

[5] For reasons of time, not every sequence was exhaustively annotated. In most cases the annotations are limited to a subset of the actual footage and/or only label every 5th frame, which is the also the rate at which the visual analysis is usually performed.

places, drawing and pointing on the whiteboard, and manipulating objects such as an hourglass, pens, and papers.

- *S6*: A group of five people engaging in a more casual social interaction in an office room filmed by a camera mounted on a tripod. At different times they are standing, walking around, shaking hands, sitting down, and leaving and re-entering the camera's view from different directions.

- *S7*: Scene shot from a raised tripod such that the camera was facing along a long corridor. Up to four people can be seen entering and leaving rooms at different distances from the camera, meeting each other, walking past or next to one another, walking towards and past the camera etc.. The camera focus and resolution were particularly poor, and the types of movement exhibited are of a kind especially challenging to many visual tracking and appearance modelling methods.

- *S8*: Sequence filmed from a camera about 1.8m above the ground facing the main entrance to the LCE. Five different people can be seen entering and leaving the Laboratory, sometimes individually and sometimes in pairs. The scene includes people opening the door for others and one person "tailgating" behind another once the door is open, which are situations of interest in terms of access security. The entrance door also marks the outer boundary of the space in which the SPIRIT system was deployed, which means that there are far fewer available SPIRIT observations.

- *S9*: A similar scenario to that depicted in sequence S8, except that the camera is filming people entering a particular office (with the camera positioned just inside the office facing the door).

- *S10*: This sequence shows up to five people in the small reception area of the LCE. They are at various points walking around, sitting down, and interacting with the wall-mounted plasma screen (which runs a version of the LabView application that may be controlled by means of the Bat).

Figure 6.14 shows frames from some of the sequences. Matching scores are computed for each frame based on manual annotations for object position. The visual processing modules were applied every 5 frames, while SPIRIT observations are usually available at least once a second for each tracked person (depending on motion and the load placed on the SPIRIT system by events in other parts of the building).

6.5.3 Multi-hypothesis Bayesian Modality Fusion

A viable multi-modal fusion method must generate reliable results that improve upon the individual modalities, while maintaining their fidelity and uncertainty information for higher-level processing and adaptation of the fusion strategy.

The approach taken here is essentially a development of Bayesian Modality Fusion [272, 242] for multi-object tracking. It uses a Bayesian graphical network (shown in figure 6.15) to integrate information from the different sources. Discrete reliability indicator variables (R_S, R_F, R_D, R_C, and R_B) are used to model how reliable each modality is at the current time. At present each variable may take on one of the values "low", "normal", and "high". The network serves as a shared template from which individual tracking hypotheses are derived. Hypotheses are instantiated by SPIRIT observations or the blob tracking framework, thus allowing tracking of people who are not tagged with a functioning Bat device or who are not currently visible by a given camera. Other visual cues such as skin colour and face detection serve as supporting modalities. Spatial and object-specific ontologies from the world model or the region segmentation and classification methods provide contextual constraints and guide the generation of hypotheses.

Reliabilities are adapted on the basis of manually specified rules over reliability indicators, such as motion and appearance variation, and performance feedback measures, such as consistency and log-likelihood of the observations under each modality. Posterior probabilities for each hypothesis can then be computed by integrating all available information using the fusion network. The position and spatial extent of tracked people is computed by reliability-weighted interpolation of the object bounding box deduced from the SPIRIT observation (if available for the current observation) and blob tracker appearance model.

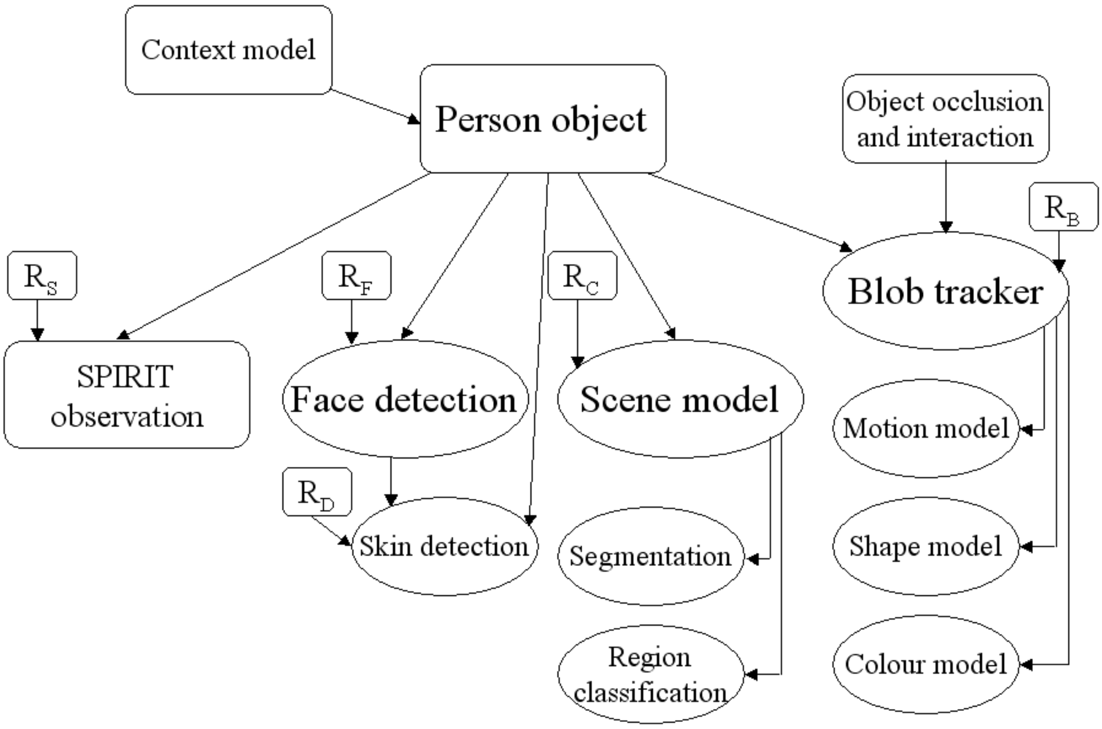


Figure 6.15: Bayesian graphical model for multi-modal fusion. Reliability variables allow adaptive integration of different sources of information.

Each hypothesis maintains its own set of reliability variables and competes for observations with other hypotheses. The conditional probabilities (including the dependency on reliability variables) of the underlying network structure were initially set by hand, but have since been re-estimated by means of the EM algorithm on statistics gathered from manually labelled training sequences consisting of over 3000 frames. Temporal evolution of the model occurs via a Kalman or particle filter applied to the colour blob tracker and through the modification of reliability variables in light of current observations. This update stage introduces a coupling of the observation models for the individual modalities. Some results from a meeting room sequence are shown in figure 6.17.

SPIRIT observations consist of projected 3D positions of Bat sensor locations

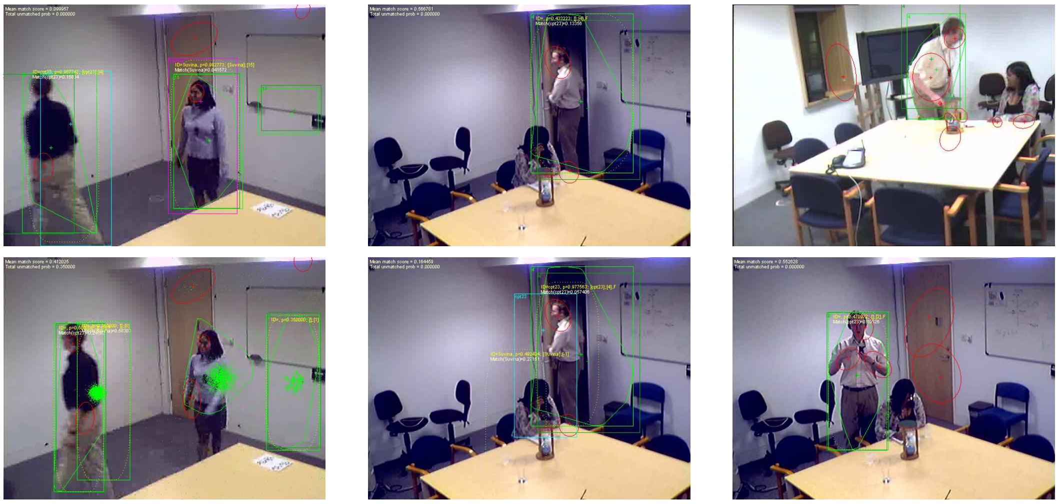


Figure 6.16: Example images illustrating the fusion process. *Top left*: **Fusing skin and face detection, Kalman filtered colour blob tracking, and SPIRIT observations.** *Bottom Left*: **Same as top left frame, except that SPIRIT observations are turned off and blob tracking is performed by the particle filter (green dots indicate blob position estimates from the particle set).** *Top middle*: **New object hypotheses can be instantiated on the basis of visual cues.** *Bottom middle*: **Integration of SPIRIT observations allows identity to be assigned, improves location estimates, and enables the tracker to identify people who are not yet identified by the visual cues alone.** *Top right* **and** *Bottom right*: **Meeting room sequence viewed from two different cameras.**

together with information on the type and identity of the observed object as available from the SPIRIT world model. The world model contains information about people's height and the probable position of their Bat on their body, and hence the projective mapping of the person's likely silhouette onto the camera's image plane can be calculated. Location events are generally quite accurate but are assigned a reduced reliability if they are not well synchronised with the current frame or if the Bat has been undergoing rapid motion.

The reliability of tracked blobs depends on the correspondence between predicted and actual position and appearance dissimilarity. The Bayesian network shown in figure 5.5 is used to reason about occlusions and object interactions.

Face detection can be a powerful cue for head position but becomes unreliable when there is too much variation in appearance due to movement, occlusions, or changes in posture. Evidence for false positives consists of detections in regions of high motion energy or areas where there is no expectation of faces being observed, i.e. where the other modalities fail to hypothesise the appearance of a person. This is particularly the case for areas of skin colour (such as a wooden door or table) where one or more faces have been detected, but which are unlikely to coincide with the appearance of a human head due to their position or shape. Conversely, face detections in regions where head presence is predicted by other modalities lead to an increase in reliability of the face cue for the given hypothesis. Skin detections may be used to model other body parts such as hands and legs. The scene model serves to disambiguate skin detections by dictating low reliability in regions that are likely to lead to false detection, e.g. wood. The scene model consists of the static region segmentation of each frame and the neural network classifications of each region. Areas of high motion energy lead to blur which degrades the reliability of the segmentation and classification. Segmentation is also unreliable when similarly coloured objects overlap.

The integration process computes a probability for each tracked object given the current observations and reliabilities. Person position is computed by weighted interpolation of the object bounding box deduced from the SPIRIT observation and blob tracker object currently associated with a given hypothesis, taking into account their respective reliabilities. Skin colour, face detections, and the scene model serve as supporting modalities, whereas the output of the SPIRIT and blob tracker maintain object identity and can serve as instantiating modalities, i.e. a new object hypothesis must be supported by either a tracked colour blob or a SPIRIT observation (see figure 6.16). In the latter case both location events for people and devices assigned to a particular person can be used. People generally wear their Bat sensor at a particular calibrated position on their body, and together with the known distance and projective properties of the camera this can be used to instantiate an expected 2D occupancy region for the person, even if no useful blob tracker can be assigned to the hypothesis at present. Face detections contribute to the combined probability if they occur within the upper third of the bounding box, and skin colour contributes if it is found anywhere within this region. Objects that are tracked only on the basis of SPIRIT information but don't appear to be visible in the current frame continue (for a while) to be represented by an hypothesis whose likelihood is adjusted according to the occlusion reasoning described above.

Hypotheses can be viewed as competing for observations, since each visual cue and SPIRIT location event may only be used to support one tracked object in a given frame. Hypotheses are removed after their probability has dropped below a threshold for a certain number of frames, but may be revived if a new SPIRIT location event tagged with the same object ID occurs. New hypotheses are instantiated in response to blob or Bat observations. Due to the relative brittleness of the visual cues alone, new hypotheses are given a low initial probability until they have been "confirmed" by a SPIRIT observation or have proven stable over several frames. This allows people who are not wearing a functioning Bat device to be tracked. SPIRIT data is also particularly valuable in maintaining object identity across occlusions (although in some cases this is also possible on the basis of blob appearance and predicted motion), and to generate expectations (expressed as hypotheses) for people who are about to enter the visual field of a camera. Hence, the Bat system and the federation of visual cues may each serve to guide the combined system's focus of attention by instantiating hypotheses and generating expectations.

6.6 Modelling of the Dynamic Environment

Having visual information as an additional sensory modality is useful when the SPIRIT system has trouble detecting a person (e.g. they are not wearing a Bat or it is temporarily concealed), or when an application requires additional information about a person's posture, direction of gaze, gestures, interactions with devices and other people, or facial expression to enhance visually mediated human computer interaction and provide a richer model of the context in which such interactions take place. In order to improve the world model's representation of personnel, the following problems can be addressed by fusing SPIRIT and computer vision information.

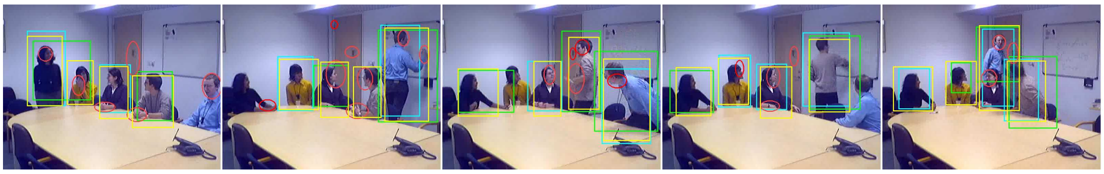


Figure 6.17: Examples of tracking results obtained for sequence S5: Rectangles denote object hypotheses derived from the two modalities (green: visual tracking, blue: SPIRIT observations) and the resulting fused hypothesis (yellow). Red ellipses indicate face detections.

6.6.1 Enhanced Tracking

Using the combined tracking framework, position estimates can be made more robust and accurate. As described in section 6.5.3, this can be achieved through Bayesian multi-modal fusion. Figure 6.17 shows sample results for a meeting

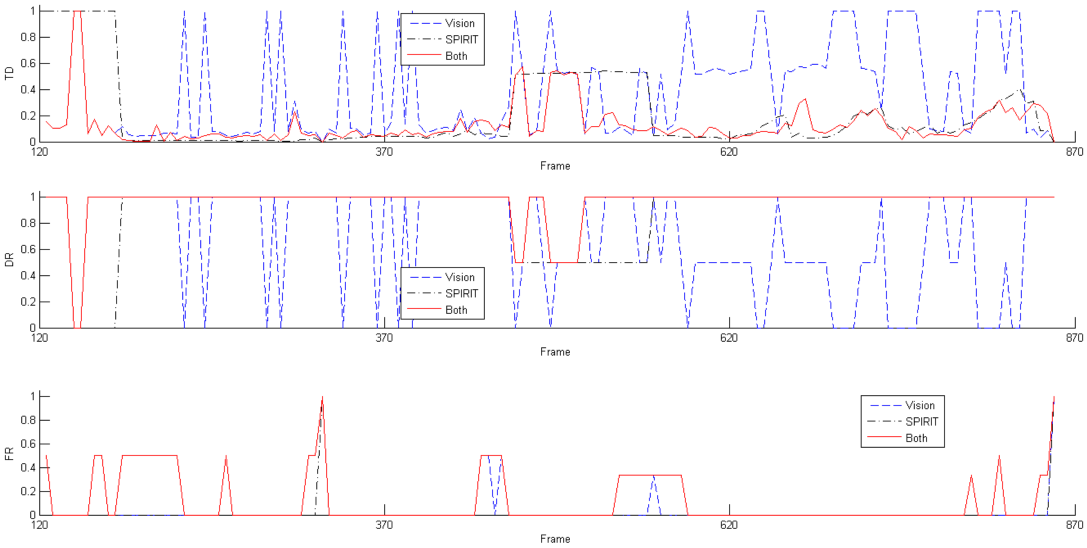


Figure 6.18: Comparative tracking results for test sequence S2 when using the two modalities, both in isolation and combined by means of the fusion process. *Top:* mean distance-from-track TD; *Middle:* detection rate DR; *Bottom:* false positive rate FR. In each case, the solid red line shows the value of the given performance measure for the outcome of the fusion method while the blue dashed and black dotted lines indicate results when using the vision and SPIRIT modalities in isolation respectively.

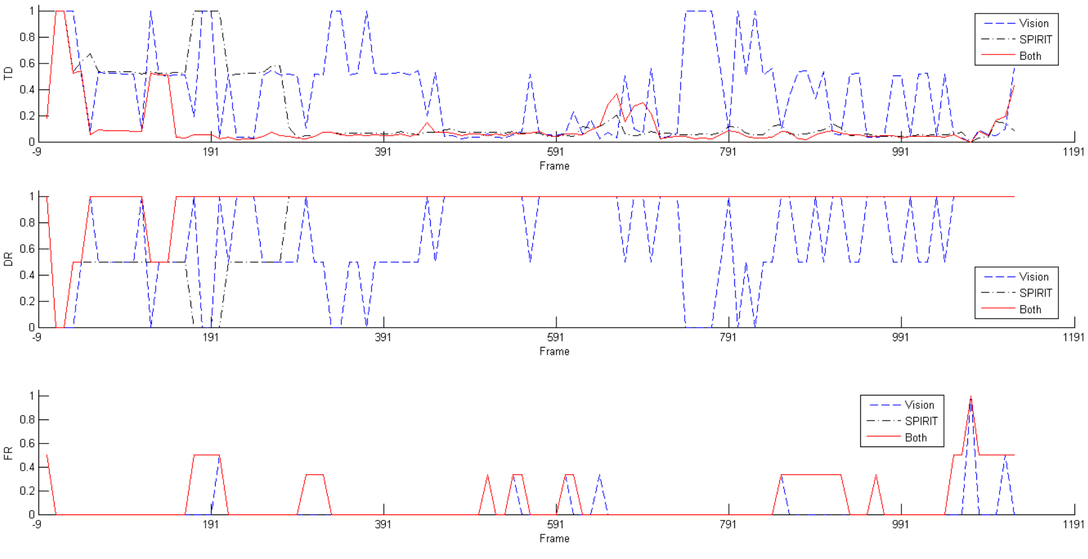


Figure 6.19: Comparative tracking results for test sequence S3.

scenario with multiple participants. As indicated below, additional information apart from a person's location can be inferred through the joint integration of the various perceptual cues.

As described in section 6.5.2, several sequences were manually annotated with

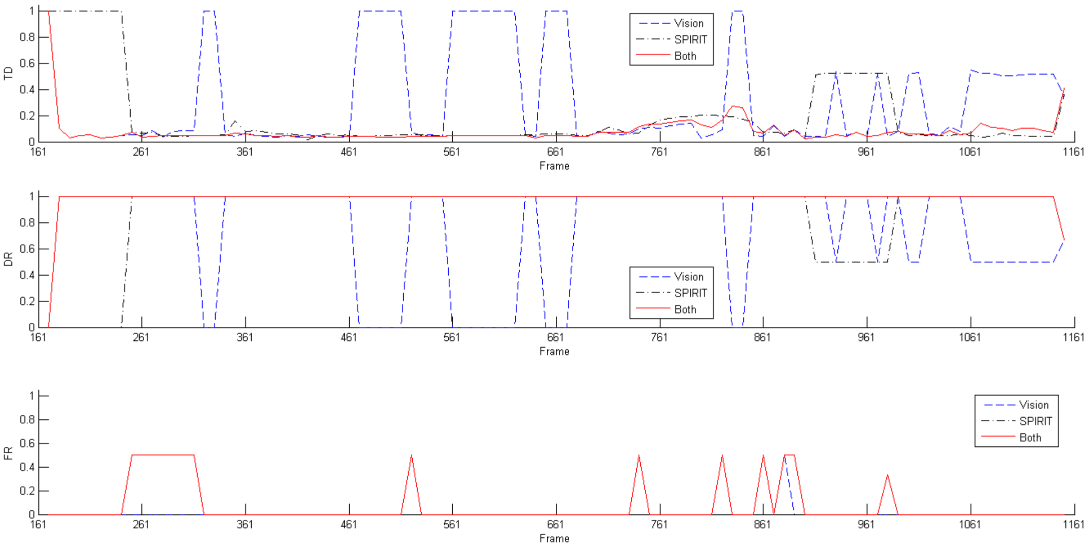


Figure 6.20: Comparative tracking results for test sequence S5.

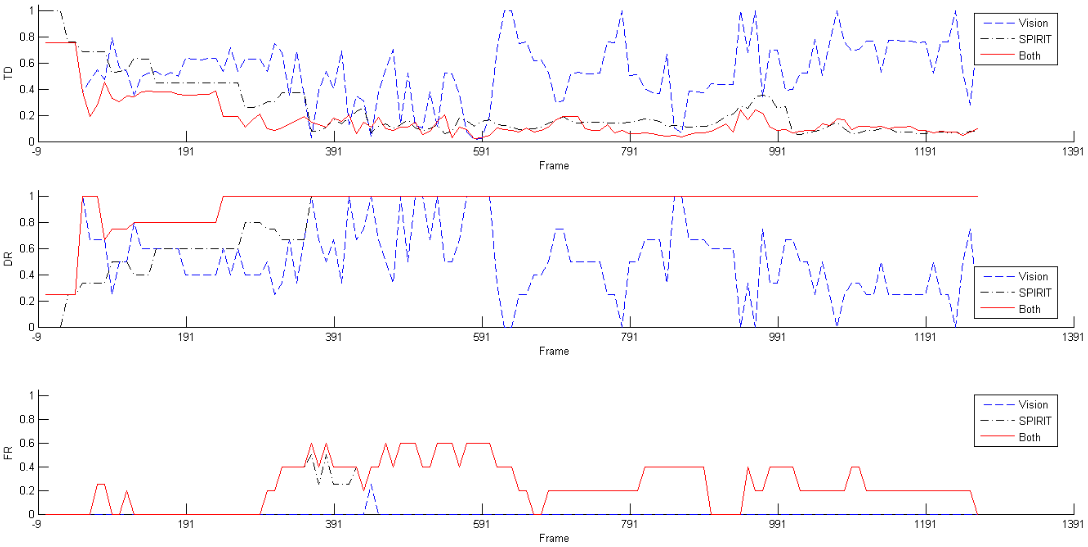


Figure 6.21: Comparative tracking results for test sequence S6.

ground truth information in order to analyse tracking performance. Figures 6.18, 6.19, 6.20 and 6.21 show performance data for sequences S2, S3, S5, and S6 respectively. For each sequence, results are shown which compare performance when tracking is performed using the two modalities on their own (i.e. only vision or only SPIRIT information) and for the fusion method described above. The performance measures plotted are the mean distance-from-track TD, the detection rate DR, and the false positive rate FR, computed for each frame in the sequence as explained in section 5.4.4. Consequently, a value of DR close to 1 indicates that all objects are being tracked in a given frame while FR close to 0 means that there are few false positives (spurious instances of objects which do not correspond to any objects marked in the ground truth for that frame). The measure TD characterises the mean accuracy of object tracks in terms of the distance between the centres of gravity of tracked objects and ground truth objects (which takes a value of 1 if no such correspondence can be established).

In order to summarise these results, figures for overall recall and precision are shown in table 6.1. As before, $Recall = \text{mean}(\text{DR})$ and $Precision = \text{mean}(N_{\text{tp}}/(N_{\text{tp}}+N_{\text{fp}}))$, where N_{tp} is the number of true positives and N_{fp} the number of false positives for each processed frame in the sequence.

As can be seen from the results, tracker accuracy and performance are generally

Modality	*Vision*		*SPIRIT*		*Fusion*	
Sequence	Recall	Precision	Recall	Precision	Recall	Precision
S2	0.674	0.907	0.857	0.976	0.963	0.852
S3	0.673	0.931	0.845	0.933	0.960	0.868
S5	0.729	0.948	0.875	1.000	0.987	0.906
S6	0.501	0.996	0.860	0.731	0.943	0.747

Table 6.1: Overall mean recall and precision for test sequences S2, S3, S5, and S6. For each sequence, recall and precision were computed by comparing tracking results obtained using vision and SPIRIT information (both in isolation and combined through Bayesian fusion) with manually labelled ground truth annotations.

enhanced by the combined fusion process compared to the outcome using vision modules or only SPIRIT observations, especially in difficult situations such as object-object occlusions. The system can exploit the multi-modal redundancies to successfully track objects that are only detected by one of the tracking components. Reliability indicators allow the system to discount modalities that fail in particular circumstances and rely on those which are likely to give accurate results, thus ensuring that the fusion process delivers results that are as good and sometimes better than those of the modality which performs best at a given time.

However, these results show that the fusion method sometimes incurs in a slightly increased false positive rate compared to either of the component modalities, which may lead to a reduction in precision alongside significant increases in recall. It is possible to optimise the precision-recall trade-off to best suit the requirements of particular applications of the sentient computing environment. For example, security applications are likely to require high recall whereas tracking for human computer interaction may need high precision.

To analyse how this can be done, some of the parameters affecting the fusion process were optimised with the goal of assessing the level of recall achievable by

Sequence	Recall	Precision	p_thresh	p_retention	fobj_retention
S2	0.979	0.974	0.75	0.20	10
S3	0.956	0.940	0.70	0.20	5
S5	0.882	0.992	0.75	0.20	10
S6	0.952	0.730	0.40	0.30	5

Table 6.2: Recall achieved by the fusion method at a comparable level of precision as that yielded by the SPIRIT modality on its own as shown in table 6.1. The values of the three variables affecting the fusion process that were optimised to achieve these results are also shown.

the fusion method at the same level of precision as that of the SPIRIT system as shown in table 6.1. No modifications were made to the Bayesian fusion network or any other aspects of the system, but 20 different combinations of values were evaluated for the three internal software variables "p_thresh", "p_retention", and "fobj_retention". In the case of results shown in table 6.1, these variables had been set to 0.40, 0.30, and 10 respectively. The parameter "p_thresh" specifies the minimum probability (as calculated by the Bayesian network in figure 6.15) that a hypothesis must initially satisfy in order to be regarded as a tracked object. In order to enable tracking of objects whose hypotheses have temporarily dropped below this threshold, "p_retention" specifies the minimum probability which must be associated with a tracked object hypothesis in order for it to continue to be tracked. "fobj_retention" specifies the maximum number of frames during which a hypothesis may be retained in this way before it is discarded, unless its probability once again rises above "p_thresh".

Table 6.2 shows resulting recall and precision values together with the values of the three aforementioned variables that were selected to bring the fusion system's precision as close as possible to that of the SPIRIT modality for each of the four test sequences. It can be seen that the fusion method exhibits greater accuracy (as measured by recall) than the Bat system for comparable levels of precision.

The current system has been implemented in Matlab and does not yet run in real-time, although it is only about 5 to 10 times slower (depending upon the type of appearance modelling employed) and could be speeded up through an optimised implementation in a compiled language. Interesting results have already emerged from offline processing to prototype potential applications of the fusion methods, as described below.

6.6.2 Appearance Modelling and Visualisation

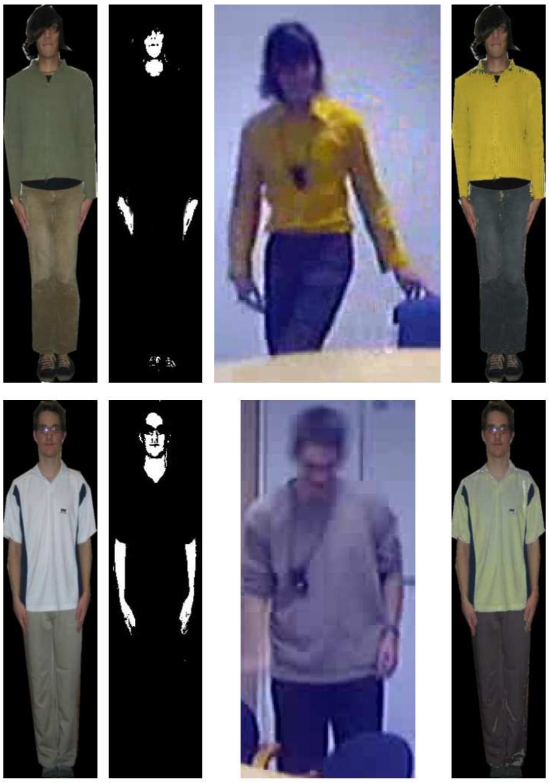
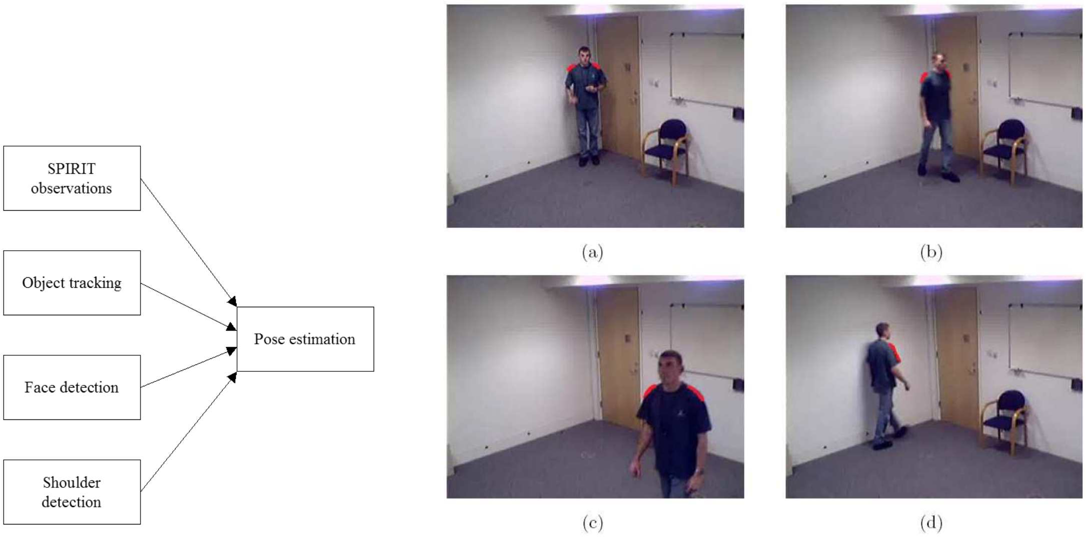


Figure 6.22: Example of dynamically updated visual appearance texture maps for two people. From left to right: Original template image used for texture mapping; Skin mask (detected skin pixels indicated in white); Sample frame showing the same person being detected by the tracking framework; Updated template resulting from a combination of the original image with colour information acquired by the colour appearance modelling.

Appearance models acquired during visual tracking are used to improve the system's visualisation capabilities. Apart from parameters such as people's height, the world model also stores a template image of each person's appearance (full frontal and back), which was taken using a digital camera. This template is used as a texture map applied to a generic 3D model to render a more faithful representation of each person in the 3D view created by the LabView application. Using the multi-modal framework, it is possible to update the colour distribution of the person's upper and lower body clothing to better represent their appearance on a given day.

This is done by performing a weighted combination of the hue, saturation, and luminance values of each of the pixels of the stored template image with the mean colour of the corresponding region on the person's current colour histogram appearance model. The process is performed separately for the upper body and leg regions. In the case of the template image, the rough boundaries between these two regions and the head are known *a priori* and skin detection is applied to prevent pixels belonging to exposed skin regions such as the hands and face from being updated. For the appearance model acquired dynamically during tracking, face and skin detection combined with the estimated height of the person (as computed from the Bat coordinates) are used to identify non-skin areas in the leg and upper body regions to determine their mean colour. Depending on the precise settings of the visual appearance modelling process, the resulting colour estimate already features robustness to noise and short-term variation. The results of the occlusion reasoning process and SPIRIT z-value are taken into account to determine whether the estimate is reliable, i.e. whether one of the two body regions is occluded dynamically (i.e. by another person) or statically (e.g. if the person is sitting down at a table).

The visual appearance of clothing is notoriously difficult to model realistically let alone acquire automatically from image data. The simple mean colour based scheme used here gave good results for its intended application, i.e. the acquisition of updated texture maps of a person's appearance through visual tracking at "run time". Different weights for the combination of hue, saturation and luminance values were tested. In order to retain the characteristic texture of the person's clothing while updating its colour, the luminance values from the template image are weighted more strongly while the weights associated with the new mean hue and saturation are set to higher values. Pixels with very high or very low luminance values must be weighted differently to ensure adequate results. The results are necessarily better for reasonably uniformly coloured garments, although the region segmentation could in principle be used to provide finer granularity of the acquired models. Figure 6.22 shows results obtained for two people tracked in sequence S5. The process is complicated by the fact that the cameras used have poor colour balance and that the lighting conditions tend to be poor. To compensate for these effects, global histogram equalisation was applied to the frames prior to estimation of the appearance models and a room-specific colour bias was used to improve the perceptual quality and representativeness of the obtained colour mean values used during updating.



(a) (b) (c) (d)

Figure 6.23: ***Left*****: Overview of the pose estimation process.** ***Right*****: Examples of successful shoulder detections.**

6.6.3 Pose Estimation

By using simple bounding box ratio tests and information from the skin and face detection modules, one can infer rough orientation (frontal, back, left and right profile view), provided the person is unoccluded at the time. While such simple estimation of basic poses and types of movement (standing, sitting, walking, bending down) is possible based on simple measures of the person's observed height and motion pattern, more sophisticated pose estimation is necessary to provide accurate orientation estimates required by some applications. For example, SPIRIT applications which dynamically project information onto a display screen or activate a device (such as a phone) in response to some event (such as Bat button being pressed or a phone call being re-routed) need information about both the target person's position and the direction they are facing in order to select the most appropriate resource accessible and available to them. A basic estimate of orientation is calculated by the SPIRIT system based on the fact that the human body shields some of the ultrasound emitted from the Bat device. However, this has proven to be inaccurate and unreliable, particularly in some locations such as the corners of rooms or in the presence of surfaces that may reflect some of the ultrasonic pulse.

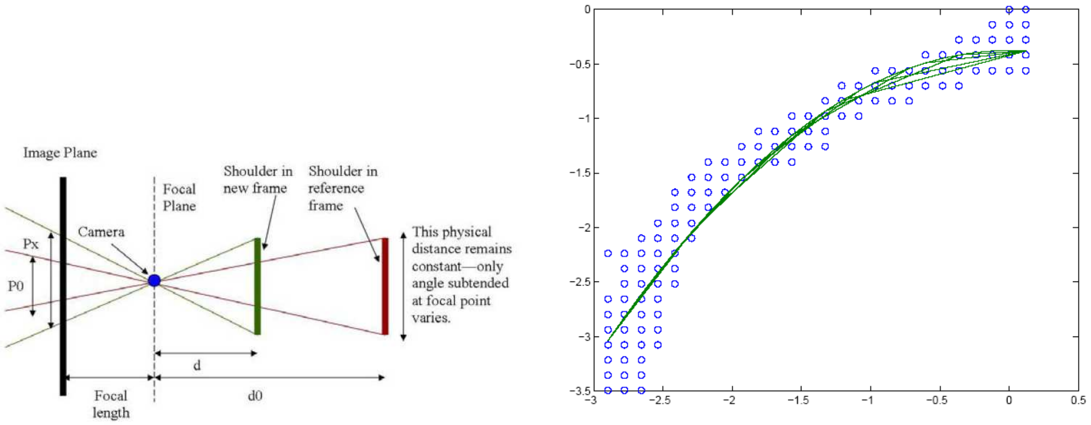


Figure 6.24: Perspective two-point algorithm for pose estimation. *Left:* **Basis of the two-point algorithm.** *Right:* **Fitting of shoulder contour to edge points.**

Better pose estimates can be computed by combining visual information with the more precise measurements of a person's location obtainable through fusion with the SPIRIT system. Two relatively simple methods were implemented for this purpose. They both rely on shoulder detection and integration of cues from visual tracking, face detection, and the Bat system observations (see figure 6.23). Detection of the shoulders is greatly simplified due to the constraints that can be inferred from the various modalities. The bounding boxes derived from the visual

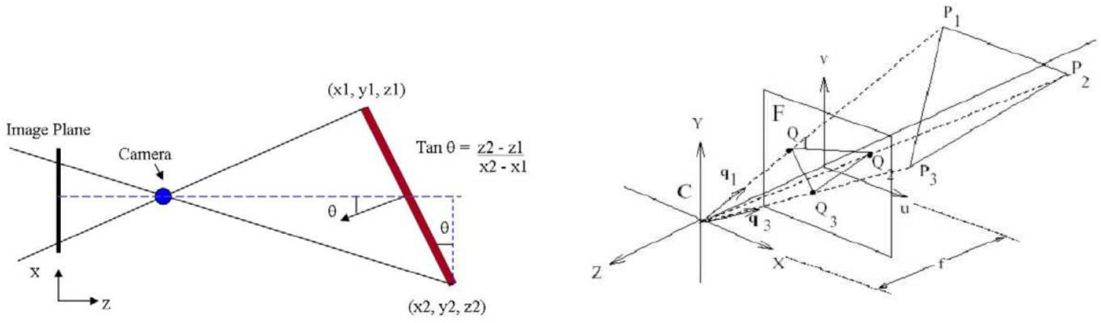


Figure 6.25: Perspective three-point algorithm for pose estimation. *Left:* **Basis of the three-point algorithm.** *Right:* **The basic geometric constraint used in the 3 point algorithm relates the distance between points P_i their projections Q_i onto the imaging plane (diagram reproduced from [239]).**

and SPIRIT based tracking hypotheses allow a good initial estimate of shoulder position. Additional inferences regarding pose can be made on the basis of face and skin detections and other contextual information. For example, whether the person is sitting down or standing up can be inferred on the basis of the z-coordinate of the person's Bat, and through occlusion reasoning combined with an analysis (by means of the neural network region classifiers) of any occluding regions (such as a wooden table or desk).

Shoulder detection is then performed by analysing the edges determined by the segmentation method within the predicted regions. Curve fitting using a second order polynomial is applied to pixels on candidate edges as shown in figure 6.24. The edge which matches (and produces the best match of any other edges) the expected shape of a shoulder contour as determined by the root mean squared error criterion is then selected within each of the two predicted shoulder regions for each person. In this way, 0, 1 or both shoulders may be identified. The midpoints of the detected shoulder curve together with the 2D position of the Bat can then serve as input to pose estimation algorithms.

There are closed form solutions to the pose estimation problem given n points [213], although the solutions are not generally unique for 3 points or fewer. In the present case one has some knowledge of the position of the Bat on the person (assuming it is always worn in roughly the same location, such as hanging from a cord around the neck), and the actual separation of the person's shoulders (either by prior measurement or by assuming a standard value). Thus, given the 3D position of the Bat (point P_3) and knowledge of the distances between the shoulder points (P_1 and P_2) and between the shoulder points and the Bat, one can infer the orientation of the plane spanned by (P_1, P_2, P_3) from the projections of these points (Q_1, Q_2, Q_3) onto the image plane (see figure 6.25). The perspective 3-point algorithm [239] assumes $P_i = a_i\mathbf{q_i}$ (where a_i is a positive real scalar and $\mathbf{q_i}$ is a unit vector) and recovers pose by solving for the parameters a_i. Since the distances $d_{ij} = |P_i - P_j| = |a_i\mathbf{q_i} - a_j\mathbf{q_j}|$ are known, one can write

$$\begin{aligned} 0 &= a_1^2 - 2a_1a_2(\mathbf{q_1}.\mathbf{q_2}) + a_2^2 - d_{12}^2 \qquad (6.26)\\ 0 &= a_1^2 - 2a_1a_3(\mathbf{q_1}.\mathbf{q_3}) + a_3^2 - d_{13}^2 \\ 0 &= a_2^2 - 2a_2a_3(\mathbf{q_2}.\mathbf{q_3}) + a_3^2 - d_{23}^2 \end{aligned}$$

This can be solved numerically by iterating the equation

$$\mathbf{A^{k+1}} = \mathbf{A^k} - \mathbf{J^{-1}A^k\,f(A^k)} \qquad (6.27)$$

where $\mathbf{A^k}$ is the vector of solutions (a_1^k, a_2^k, a_3^k) at the kth iteration, $\mathbf{f}(\mathbf{A^k})$ is the current vector of solutions to the set of equations 6.26 given parameters $\mathbf{A^k}$, and $\mathbf{J}$ is the Jacobean matrix of partial derivatives of $\mathbf{f}$ with respect to the current parameter values. The pose angle Θ can then be calculated from the estimated 3D shoulder coordinates as shown in figure 6.25.

However, the 3-point algorithm was found to be sensitive to errors in the measurements of the shoulder points and Bat location. One can often make the simplifying assumption that the 3 points (two shoulder points and one Bat position) lie on a plane that is orthogonal to the ground plane, i.e. the person is upright. In that case the 3D Bat position can be used to estimate the distance from the shoulder points to the camera and the pose can now be estimated from the shoulder points alone. The 2-point algorithm thus simplifies to computing the pose angle from

$$\Theta = \cos^{-1}(\frac{di_{12}}{dn})$$

where di_{12} is the current distance between the shoulder points in the image and *dn* is their projected separation if the person were facing the camera (see figure 6.24).

The 2-point algorithm typically gives more accurate and more robust results in practice on the testing sequences. In order to better deal with errors caused by the pose estimation algorithms, one can impose a motion continuity constraint that filters out discontinuous changes in pose. This also allows an estimate of the direction a person is facing to be made when only one of the shoulders is visible. The approach could be extended using a model-based pose estimation method such as [155].

6.6.4 Identification and Verification

Visual information can be of use in detecting potential intruders (people not wearing functioning Bats) and providing vision-based biometric verification of identity.

By comparing the number and identity of people tracked visually and by the SPIRIT system, people who are not wearing a functioning Bat can be detected. For security purposes, one can easily envisage an application which raises an event if such sightings are not expected to occur and forwards an image of the person(s) in question to an administrator.

Another application currently under development concerns user authentication for security critical applications, e.g. those which allow users to automatically unlock office doors or automatically login to a computer in their vicinity. The SPIRIT system currently provides security only on the basis of "something you have" (the user's Bat device), which may be augmented by applications also requiring "something you know" such as a password. Using computer vision, this could be enhanced by "something you are", i.e. by integrating visual biometric verification [20] of identity through analysis and recognition of facial characteristics, gait, iris patterns etc..

By utilising the combined fusion mechanisms, information from the SPIRIT system could be integrated to pose the problem in terms of verification (i.e. whether the person is who their Bat indicates they are) rather than the much harder problem of unconstrained recognition. In addition, the SPIRIT system allows such prior information about the person's purported identity and their actual biometric signature to be acquired in a non-intrusive and transparent way. The system could use the Bat sensor information to detect that a person is present and select one or more camera views to verify their identity as indicated by the identity tag of their Bat. The location of the Bat can then be used to constrain the search window for a head and face detector, which forwards an image of the detected face to a face recogniser. Methods which perform recognition on a sequence of images as opposed to a single frame might offer additional improvements in performance [94, 163, 300, 160]. Rather than solving the extremely difficult problem of general face recognition, visual authentication can then be approached as a verification problem. Moreover, it can be greatly constrained by fusing other kinds of information about assumed identity, face location, lighting conditions, and local office geometry.

6.7 Modelling of the Office Environment

The world model's static component can also be augmented using vision-based environmental modelling techniques. In particular, solutions to the following tasks have been implemented.

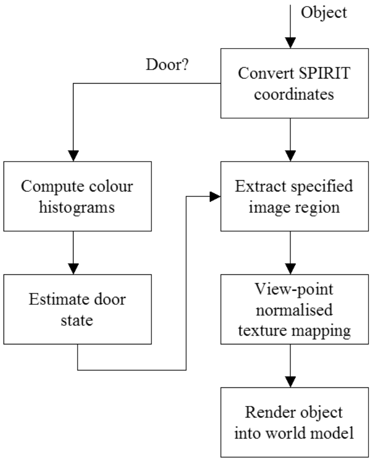

Object
Door?
Convert SPIRIT coordinates
Compute colour histograms
Extract specified image region
Estimate door state
View-point normalised texture mapping
Render object into world model

Figure 6.26: Overview of the object extraction and state estimation process.

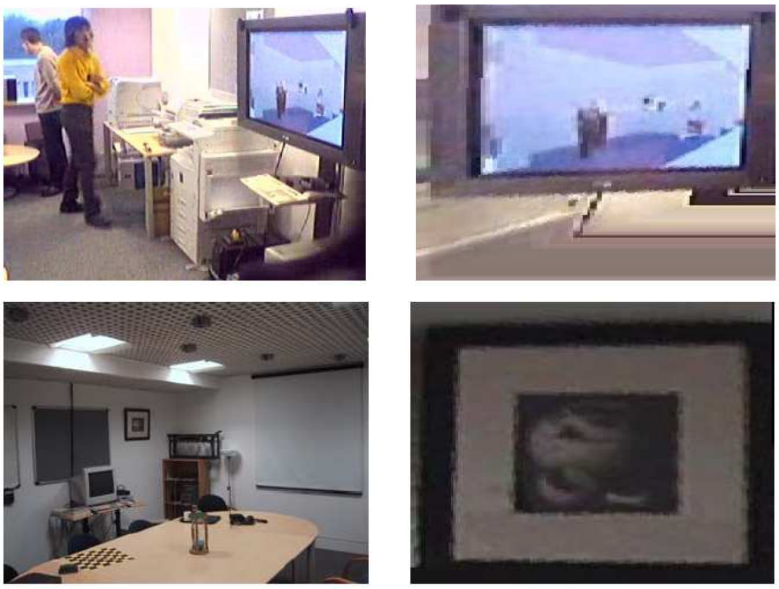


Figure 6.27: Perspective correction and texture mapping examples. *Top:* **Plasma screen.** *Bottom:* **Picture frame.**

6.7.1 Enhanced Visualisation

The visualisation of rooms can be made more compelling by acquiring texture maps of actual features such as decorations, carpet colouration, and white board surfaces. Figure 6.26 presents an overview of the steps involved in modelling textured surfaces and inferring dynamic state (door open or closed, see section 6.7.2) by integrating visual and SPIRIT information.

Given that the view-angle and position of a camera can be calibrated and correlated with the reference frame and office metadata of the world model, it is possible to reverse the projective mapping introduced by a camera and acquire viewpoint normalised images of objects and surfaces of interest which can be used for texture mapping. Such texture maps are then used to provide enhanced and up-to-date visual models of the environment that can be viewed using applications such as LabView as discussed above.

Figure 6.27 shows two frames from test sequences filmed in the sentient office. Regions of interest (a plasma screen and a picture respectively) were selected. Their 3D corner points in the SPIRIT world model were then converted to the corresponding 2D coordinates in the image acquired by the camera by means of the

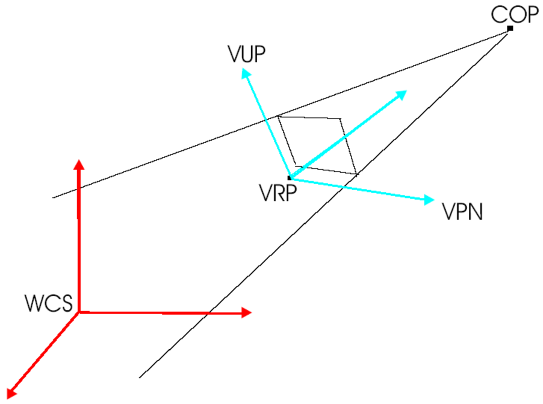


Figure 6.28: Perspective correction method for view normalised texture mapping of real world objects and surfaces. The camera reference frame is defined in the world coordinate system (WCS) through the view reference point (VRP), the view plane normal (VPN), the view up vector (VUP), and the centre of projection (COP).

calibrated camera parameters. Since the orientation of the original surfaces in the 3D SPIRIT world frame with respect to the camera is known from the coordinates, the distortion caused by the projection onto the imaging plane can be reversed in order to obtain a view normalised texture map of the given object in the original scene. Figure 6.28 illustrates the geometric relationships between the SPIRIT and image plane coordinate systems. Once the relationship between camera and world (i.e. SPIRIT) coordinate frames has been established, a normalised (full frontal) view of the object or surface in question can be recovered numerically by inverting the perspective relationship. An example of the texture map of a door acquired in this way is given in figure 6.29.

Such texture maps can then be rendered as part of the LabView application to augment its 3D representation of the environment. Figure 6.30 shows an example of such enhanced room visualisation. This would be of use in a range of applications such as video conferencing and seamless shared virtual environments. Participants located in different physical offices equipped with the sentient computing system could thus access and browse a unified view onto the world model, which includes dynamic features of interest such as the location and identity of people and the visual content of resources such as certain displays, whiteboards, and projector screens.

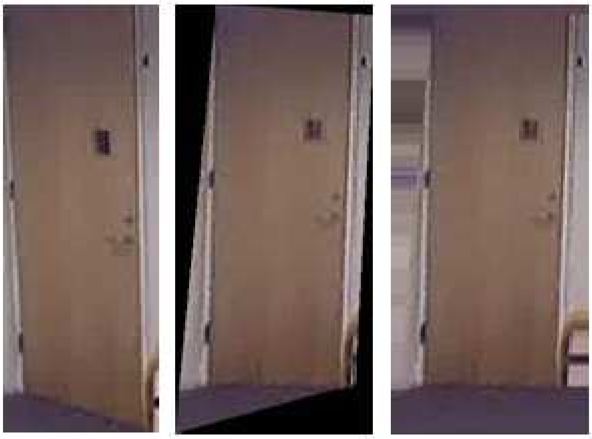


Figure 6.29: Example of the perspective correction applied to the image of a door. Given 3D coordinates of the door frame in the SPIRIT world model, the corresponding 2D coordinates of the door image projected onto a given camera are computed from the calibration information. The portion of the door that is currently visible is then cropped, rotated by the appropriate angle, and re-normalised (which may require some extrapolation of pixels along the boundary) to derive a rectangular image which can be used as a texture bitmap for visualisation purposes.

6.7.2 State Estimation

Aspects of environmental state such as that of doors (open, closed, ajar) and lights (on, off) can be determined.

Door and light state can be estimated quite robustly using simple colour histogram thresholding techniques. The classifier is a simple Support Vector machine that was trained on PCA transformed images of the door region. Figure 6.31 shows examples of some of the images used in training of the door state detection module and examples of correct classifications. Although some types of occlusions cause mis-classifications, the method currently correctly identifies door state in over 80% of cases on a challenging test set.

6.7.3 Environmental Discovery

Semi-static parts of the environment such as furniture and computer screens can be recognised using vision techniques.

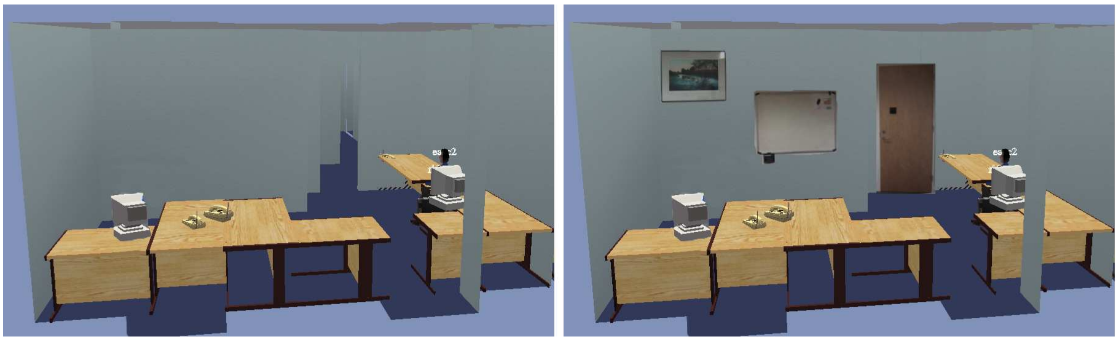


Figure 6.30: Visually augmented representation of the world model. *Left:* **LabView generated view of an office.** *Right:* **The rendered scene has been augmented with texture maps acquired from a calibrated camera.**

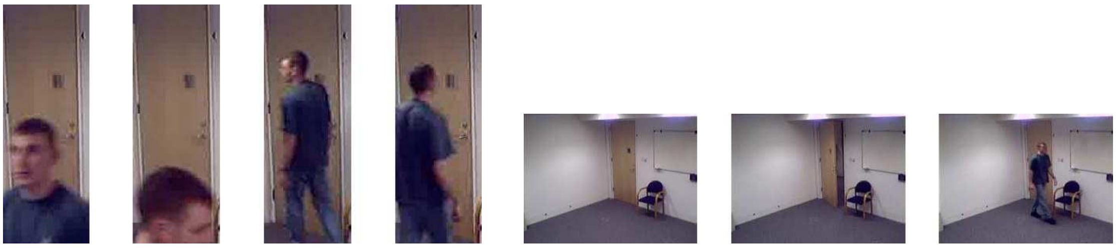


Figure 6.31: Door status detection. *Left:* **Example images used for training of the door state classifier for state "closed".** *Right:* **Examples of correct classifications ("closed", "open", "closed").**

As shown in figure 6.32, neural networks were trained to classify segmented images according to categories of man-made material (wall, carpet, cloth, and wood). As shown in table 6.3, the classifiers achieve correct classification scores of up to 96.5% on a large (about 500 images) test set of images taken in the LCE

Category	*Classifier type*	Classification score
Wood	Multi-layer perceptron	92.8%
Wall	Radial basis function network	93.0%
Cloth	Radial basis function network	95.9%
Carpet	Radial basis function network	96.5%

Table 6.3: Performance of neural network image region classifiers for man-made materials.

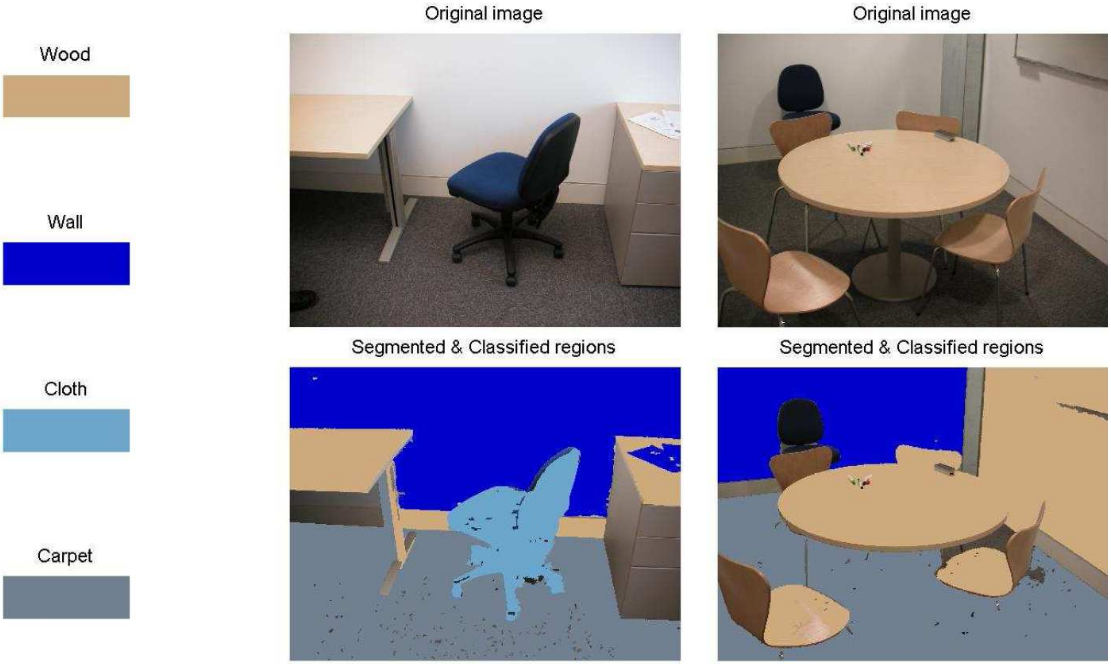

Wood
Wall
Cloth
Carpet
Original image
Original image
Segmented & Classified regions
Segmented & Classified regions

Figure 6.32: Examples of neural network based classification of man-made material categories in segmented images of the sentient office.

offices. Figure 6.34 shows the Receiver Operating Characteristic (true positive versus false positive rate) curves for the classifiers.

Classified image regions can then serve as an intermediate representation for object detection and recognition. Figure 6.33 shows two examples of Bayesian networks which were trained to detect two different categories of chairs commonly

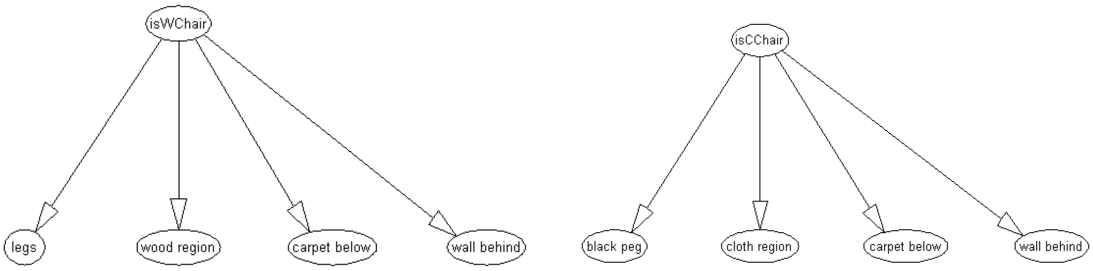
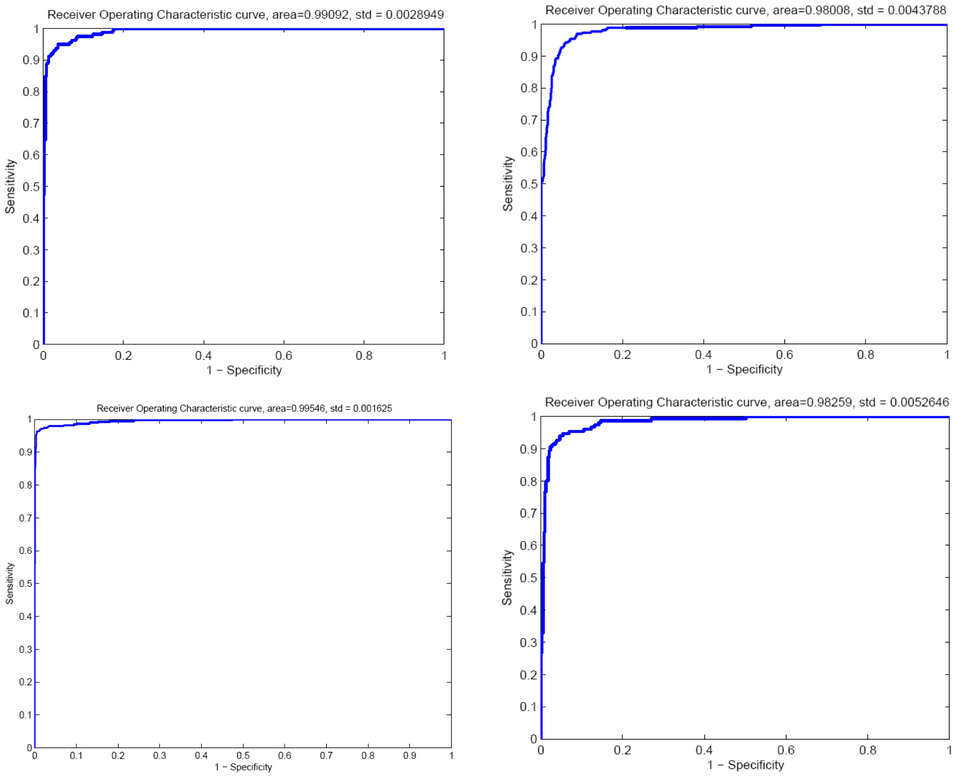


Figure 6.33: Bayesian networks for recognition of wooden (left) and cloth-backed (right) chairs based on image segmentation and classification information.



Receiver Operating Characteristic curve, area=0.99092, std = 0.0028949

Receiver Operating Characteristic curve, area=0.98008, std = 0.0043788

Receiver Operating Characteristic curve, area=0.99546, std = 0.001625

Receiver Operating Characteristic curve, area=0.98259, std = 0.0052646

Figure 6.34: Receiver Operating Characteristic (ROC) curves for the neural net classifiers for (left to right, top to bottom): wood, wall, cloth, and carpet.

found in the office where the SPIRIT system has been deployed. The detectors are fairly robust with respect to scale, position and orientation of the target objects, although occlusions remain a challenge. Figure 6.35 shows some examples of detections for three different classes of objects.

There is scope for combining visual information with other techniques for dynamic environmental discovery such as a statistical analysis of ray traces from the ultrasonic pulses emitted by people and autonomous robots [103].

6.8 Summary

As computer vision continues to mature, it is likely to play an increasingly important role in the rapidly growing field of ubiquitous computing. This chapter presents a novel approach to harnessing the benefits of computer vision within the

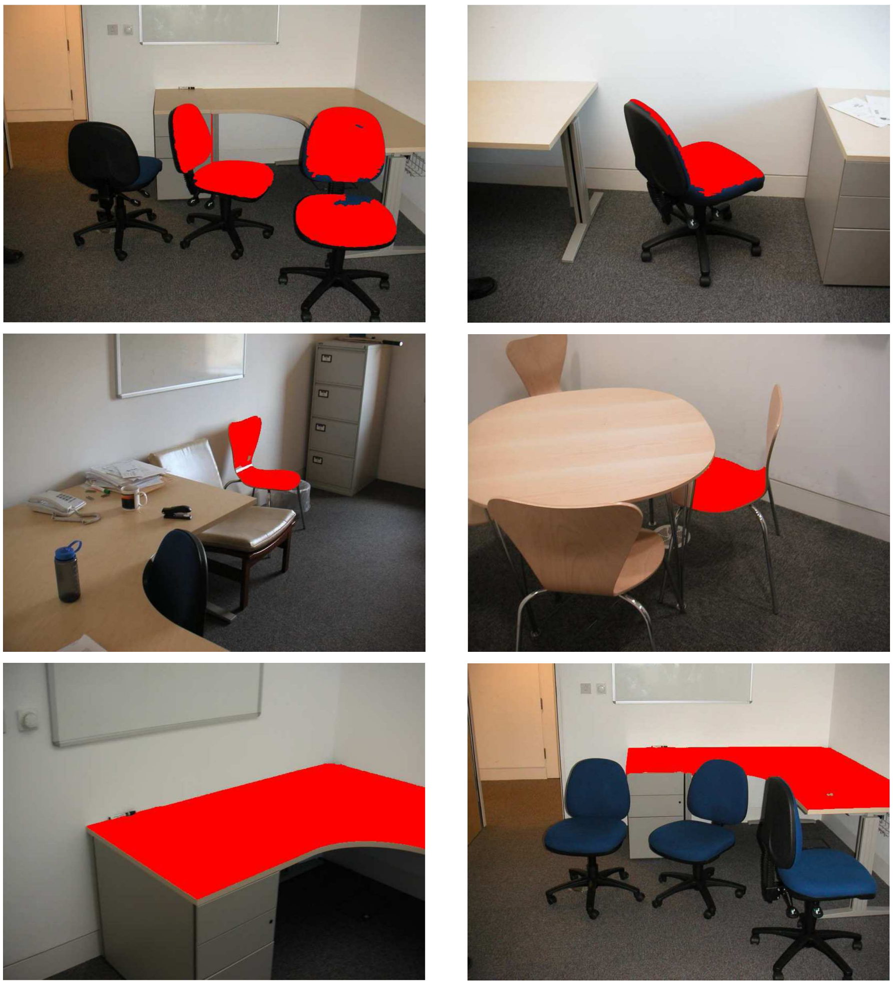


Figure 6.35: Examples of object detection results obtained by the Bayesian networks (detected objects are marked in red). Top: cloth-backed chairs; Middle: wooden chairs; Bottom: desks.

context of a sentient computing system deployed throughout an office space. It is shown how different computer vision methods such as tracking algorithms and appearance models can be fused with information from an ultrasonic tracking system to significantly augment the capabilities and robustness of the system's world model.

The approach is founded on a Bayesian framework for adaptive fusion of different sources of information. It uses an ontology of object and environmental properties to integrate different hypotheses about the perceived context. The world model serves as a shared representation of context that is made available to users and applications. Knowledge gained from a range of computer vision methods has proven effective in enhancing the world model, thus allowing the sentient computing system to maintain a richer and hence more useful representation of the environment and events that occur within it. The sensor fusion and information integration adds value to both the visual and ultrasonic modality by complementing their capabilities and adapting to error characteristics exhibited by the different sources of information at different times.

The sentient computing system provides a variable granularity spatial model of the environment and a reliable device tracking facility that can be used to automatically (re)initialise and re-focus vision modules whenever an event or scene context of interest is observed by a camera. A number of applications of the sentient computing technology can in turn benefit from the video interpretation framework through the fusion of the ultrasonic and visual modalities.

The envisaged usage is very much in line with the proposed framework that integrates multiple sources of perceptual information in order to derive an internal queryable representation suited to a particular application context. By ensuring that the symbolic inferences drawn by the system remain grounded in the signal domain, the system can support a range of possible queries as inferences and adapt its hypotheses in light of new evidence.

To ensure sufficient performance to enable real-time processing, the fusion of individual perceptual modalities can be set up as a hierarchy where inexpensive detectors (e.g. finding the rough outline of a person) narrow down the search space to which more specific modules (e.g. a face spotter or gesture recogniser) are applied. The system thereby remains robust to error rates by integrating information vertically (applying detectors with high false acceptance rates to guide those with potentially high false rejection rates) and horizontally (fusing different kinds of information at the same level to offset different error characteristics for disambiguation). In this way, vision is used to enhance the perceptual inference capabilities of the sentient computing infrastructure by adding further sources of information to update, query, and extend the system's internal ontology and external event model.

By maintaining a notion of its own internal state and goals the system can restrict its focus of attention to perform only those inferences which are required for the current task (e.g. verifying the identity of a person who just entered the visual field). Real-time requirements and other resource limitations could be used as additional constraints for the fusion process. Since all the information gathered by vision is integrated into the sentient computing system's world model, additional concerns such as user privacy can be addressed through existing or proposed mechanisms that are applied to the derived context information [259, 21].

Chapter 7

Conclusions

7.1 Discussion and Summary of Achievements

Ontologies are an important tool in many branches of science, technology, philosophy and linguistics. They offer a convenient means of encoding hierarchical knowledge in terms of entities, attributes and relationships that may be used to characterise a given domain. Ontology is the theory of objects in terms of the criteria which allow one to distinguish between different types of objects and the relations, dependencies, and properties through which they may be described. Ontologies encode the relational structure of concepts which one can use to describe and reason about aspects of the world. They can play both a passive representational or taxonomic role in terms of prior ontological commitments to certain objects and categories ("that which exists in the world", "that which we are looking for") and an active inferential role which defines how the relevant properties of the world may be identified ("how to find what we are looking for"). This makes them eminently suitable to many problems in computer vision that require prior knowledge to be modelled and utilised in both a descriptive and prescriptive capacity.

However, one must beware of the *inductive bias* that can cause people to read too much into a given ontology. For example, simply having terms such as "run", "fight", etc. in the ontology does not mean that these concepts will be sufficiently well defined for the purposes of a recognition system built upon it. In order to be of practical use, an ontology must therefore be grounded in reality, i.e. the data accessible to the system. This requires mechanisms for representing and assessing the *individuation criteria* [38, 248] by which objects and their properties are differentiated according to the ontology.

This thesis 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. The efficacy of the proposed approach is demonstrated through the development and analysis of solutions to a range of challenging visual analysis tasks.

As noted above, 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 that 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 thesis.

Furthermore, the notion of ontology based languages is introduced as a powerful means of creating computational vehicles for knowledge representation and matching that 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.

To illustrate and validate these approaches, three distinct problem domains in computer vision are considered in this thesis:

- Chapter 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 *prescriptive* ontology of image content descriptors, which is hierarchically decomposed using a language that 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 that exist in the ontology, the OQUEL queries, and the content of images provide a basis of inference and contextual disambiguation through which the ontological language can be extended with new terms.

- In chapter 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 chapter 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 that improves the performance of the Bayesian networks. As in the preceding chapter, 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.

- The final problem area presented in chapter 6 concerns the fusion of multiple sensory modalities to realise the vision of *sentient computing.* The proposed methods integrate information inferred through a range of computer vision tracking and appearance modelling techniques applied to video data from calibrated cameras with that obtained from an ultrasonic location system. Fusion is carried out with reference to a world model that acts as a central ontology and shared language of context information. In this way, the system is able to maintain the illusion of a shared perception between users and applications of the sentient computing system. The system's perceptual capabilities are considerably enhanced through the fusion of visual information, while the world model provides strong prior information that is used to initialise and drive the integration process through the generation and evaluation of new hypotheses over relevant aspects of the underlying ontology.

  The world model thus serves as a *descriptive* ontology whose static components are implicitly grounded, while dynamically updatable aspects of the model are grounded through sensor systems. By sharing the same ontology, different sensor systems can be integrated without having to explicitly translate between their internal models of the world. Information obtained through the integration of computer vision techniques augments the capabilities and robustness of the world model, thereby allowing it to maintain a richer and more accurate ontology of context. In this way, the semantic gap, or *context gap*, between the applications built upon the system and their users is reduced. Applications drive the inference process by posing queries to the system, resulting in an update of the system's internal state, in response to user actions or automated events. These requests for access to context knowledge thus serve as goals and free the system from falling prey to the *frame problem*.

7.2 Avenues for Further Work

A number of promising refinements and extensions of the applications of the work presented in this thesis are conceivable, as was already touched upon in the relevant chapters.

The methods proposed for multi-sensory fusion and augmentation of the SPIRIT world model were shown to be stable and yield sufficiently high performance such that a range of applications could be made to benefit from them. This would necessitate additional software engineering effort to fully integrate them into the SPIRIT system for continuous autonomous deployment. Several interesting lines of further research follow on from this, for example the optimal integration of information from multiple cameras and multi-modal tracking across a large office space. As discussed in chapter 6, a range of new applications such as vision-based identity verification, environmental discovery, and enhanced personnel modelling would then be available to further the realisation of the sentient computing paradigm.

The ontological query language (OQUEL) proposed in chapter 4 offers great potential for further development and novel research. While current results demonstrate its power and versatility as applied to general purpose photographic image retrieval, additional processing modules and sources of information could be integrated into the framework. While the ICON system already facilitates search over camera metadata properties and textual image annotations, these could be integrated more explicitly into the OQUEL ontology. This would also facilitate customisation of OQUEL for more specialised image archives, e.g. museum collections, medical images, and art libraries, which frequently feature pre-existing text-based annotation schemes. Development of additional recognition methods for facial attributes such as expression, orientation, age, gender, etc. would also be useful in supporting higher-level queries.

The OQUEL ontology is extensible, and already features objects (e.g. "people", "animals"), relations (e.g. "close to", "similar"), attributes (e.g. "green", "circular"), and other category labels such as "sky" and "winter" (which are neither objects nor relations [38]). The choice of additional visual categories is task dependent. Elementary descriptors can be thought of as undergoing a transition from noun to adjective and thus represent a very flexible basis language for composite object recognition. Moreover, the work presented in chapter 5 could be integrated to extend OQUEL into the video domain by incorporating object states, roles, actions, scenarios and events into the ontology. As before, the query-guided nature of the inference process and the semantic and syntactic structure of queries and content descriptions would provide a powerful way of defining spatial and temporal context. Additional information such as sound and speech analysis or closed caption text parsing could also be integrated.

As will be apparent from the preceding chapters, a primary focus of this thesis are the computational structures required to model semantically labelled entities and their dual parametric representations in terms of the underlying data. The proposed framework is therefore not limited to computer vision information, as was already demonstrated in this dissertation. There is great potential in applying the methods introduced herein to domains such as language understanding and text retrieval. The possibility of augmenting the OQUEL framework with natural language parsing techniques was already touched upon. This would not only be of value in increasing the power of the query interface and enhancing the richness of the query language, but also opens avenues of further investigation. A very promising line of further research would be to extend the notion of ontological query languages and self-referential inference to general purpose reasoning and conceptualisation. Concepts in an ontological language as proposed in this thesis are defined with reference to those entities to which they pertain in the current context. Words carry explicit semantic associations, for example in OQUEL there is a testable chain of evidence from a term like “grass” or “above” to derived image properties. Ontological resources for lexical semantics such as WordNet [175] can be seen as complementary to OQUEL in that they group words according to semantic relationships. Such categorisations and word associations however fail to perform symbol grounding: nowhere in the canon of knowledge in WordNet on the term “green” is there any understanding of what it is for a real object to be green.

OQUEL and schemes derived from it could thus be seen as a precursor to more general semantic node based models for knowledge representation from images, video, text and other sources. In particular, they offer a way of defining a well-founded notion of semantic similarity, something which opens up many exciting opportunities for further research. It was already shown that visual similarity and semantic similarity are very different in the case of image retrieval, a factor that greatly contributes to the semantic gap between people and computers. Concepts expressed as sentences in an ontological language can be compared intensionally on the basis of their syntactic structure, and extensionally by evaluating them over a particular data set and analysing the resulting matchings. Conceptual matching therefore proceeds by a process of synthesis and analysis. This idea could also provide a way of translating between different ontologies and representations of concepts by comparing the way in which their internal concepts are linked to the world. Such work can be seen as complementary to the MindNet [67, 173] project. MindNet automatically acquires semantic networks of concepts based on natural language parsing and analysis of statistical word associations. Once again, the approach presented in this dissertation would offer a means of generalising MindNet to non-textual data.

By forming associations between concepts and their instantiations through ontological sentences, one could also bootstrap the process of knowledge acquisition. Such associations can be seen as *weakly labelled* data, and recent techniques in statistical supervised machine learning [223] may then be used to iteratively improve the definition of concepts and use these concepts to identify new instances in the data. This overcomes one of the conundrums of statistical learning, namely that candidate objects must be known before their statistics can be collected, while statistics are used to define suitable candidate objects in the first place. In practice, this process could be achieved by building a new version of OQUEL for text and/or multimedia data and bootstrapping its ontology through acquisition of weakly labelled data obtained from the web. As evidenced by the success of weakly indicative retrieval metrics for the internet, the web is a rich source of document content, context, co-occurrence statistics, and link relationships through which concepts can be defined extensionally. There are therefore fruitful avenues of research in exploring ways of integrating this approach with the newly proposed semantic web to enable more general semantics without a need for explicit pre-annotation.

More generally, as shown in chapter 3, the inference framework proposed in this thesis gives rise to the notion of a self-referential paradigm for perception. The central idea is to have a system with an internal queryable model of state representing a view of the world, a set of (representational) goals, and a feedback loop whereby an iterative process of introspection followed by renewed interpretation leads to an improved world model. Since ontological sentences are grounded in reality and disambiguated in a context dependent upon the current goal set, they present the prospect of a flexible substrate for high-level reasoning. By the same token they might also serve as a computational model of certain theories of human cognition.

Bibliography

[1] A. Abella. *From Imagery to Salience: Locative Expressions in Context.* PhD thesis, University of Columbia, 1995.

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

[3] R. Ackoff. From data to wisdom. *Applied Systems Analysis*, 16, 1989.

[4] M. Addlesee, R. Curwen, S. Hodges, J. Newman, P. Steggles, A. Ward, and A. Hopper. Implementing a sentient computing system. *IEEE Computer*, 34(8):50–56, 2001.

[5] M. Addlesee, A. Jones, F. Livesy, and F. Samaria. The ORL Active Floor. *IEEE Personal Communications*, 4(5), 1997.

[6] N. Adly, P. Steggles, and A. Harter. SPIRIT: A resource database for mobile users. In *Proc. of the ACM CHI'97 Workshop on Ubiquitous Computing*, 1997.

[7] M. Adoram and M. Lew. IRUS: Image retrieval using shape. In *Proc. Int. Conference on Multimedia Computing and Systems*, 1999.

[8] J. Aggarwal and Q. Cai. Human motion analysis: A review. *Computer Vision and Image Understanding: CVIU*, 73(3):428–440, 1999.

[9] A. Aho and T. Peterson. A minimum distance error correcting parser for context free languages. *SIAM Journal of Computing*, 1(4), 1972.

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

[11] P. Bahl, V. Padmanabhan, and A. Balachandran. Enhancements to the RADAR user location and tracking system. Technical report, Microsoft Research, 2000.

[12] H. Barlow. The knowledge used in vision and where it comes from. *Philosophical Transactions of the Royal Society London B*, 352:1141–1147, 1997.

[13] H. Barlow. The exploitation of regularities in the environment by the brain. *Behavioural and Brain Sciences*, 24:602–607, 2001.

[14] K. Barnard, P. Duygulu, and D. Forsyth. Clustering art. In *Proc. Conference on Computer Vision and Pattern Recognition*, 2001.

[15] K. Barnard, P. Duygulu, D. Forsyth, N. Freitas, D. Blei, and M. Jordan. Matching words and pictures. *Journal of Machine Learning Research*, 3:1107–1135, 2003.

[16] K. Barnard and D. Forsyth. Learning the semantics of words and pictures. In *Proc. International Conference on Computer Vision*, 2001.

[17] M. Beal, H. Attia, and N. Jojic. Audio-video sensor fusion with probabilistic graphical models. In *Proc. European Conference on Computer Vision*, 2002.

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

[19] H. Beck, A. Mobini, and V Kadambari. A word is worth 1000 pictures: Natural language access to digital libraries. In *Proc. Second World Wide Web Conference*, 1994.

[20] S. Bengio, C. Marcel, S. Marcel, and J. Mariethoz. Confidence measures for multimodal identity verification. *Information Fusion*, 3:267–276, 2002.

[21] A. Beresford and F. Stajano. Location privacy in pervasive computing. *IEEE Pervasive Computing*, 2(1):46–55, 2003.

[22] J. Black, T. Ellis, and P. Rosin. A novel method for video tracking performance evaluation. In *Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, 2003.

[23] A. Blake and M. Isard. *Active contours.* Springer-Verlag, 1998.

[24] A. Bobick, S. Intille, J. Davis, F. Baird, C. Pinhanez, L. Campbell, Y. Ivanov, A. Schutte, and A. Wilson. The KidsRoom: A Perceptually-Based Interactive and Immersive Story Environment. Technical Report 398, MIT Media Lab, 1996.

[25] A. Bobick and Y. Ivanov. Action recognition using probabilistic parsing. In *Proc. Conference on Computer Vision and Pattern Recognition*, 1998.

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

[27] H. Boehme, U. Braumann, A. Brakensiek, M. Krabbes, A. Corradini, and H. Gross. User localisation for visually-based human-machine-interaction. In *Proc. Int. Conference on Automatic Face- and Gesture Recognition*, 1998.

[28] J. Borges. *Borges: A Reader*, chapter The Analytical Language of John Wilkins. New York:Dutton, 1981.

[29] J.-Y. Bouguet. Matlab calibration toolbox. `http://www.vision.caltech.edu/bouguetj`.

[30] S. Brandt. Use of shape features in content based image retrieval. Master's thesis, Helsinki University of Technology, 1999.

[31] B. Brumitt, B. Meyers, J. Krumm, A. Kern, and S. Shafer. Easyliving: Technologies for intelligent environments. In *Handheld and Ubiquitous Computing*, pages 12–29, 2000.

[32] H. Bunke and D. Pasche. *Structural Pattern Analysis*, chapter Parsing multivalued strings and its application to image and waveform recognition. World Scientific Publishing, 1990.

[33] H. Buxton. Learning and understanding dynamic scene activity. In *Proc. ECCV Generative Model Based Vision Workshop*, 2002.

[34] H. Buxton and S. Gong. Advanced visual surveillance using Bayesian networks. In *Proc. International Conference on Computer Vision*, 1995.

[35] H. Buxton and N. Walker. Query based visual analysis: Spatio-temporal reasoning in computer vision. *Vision Computing*, 6(4):247–254, 1988.

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

[37] J. Canny. A computational approach to edge detection. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 8(6), 1986.

[38] B. Cantwell-Smith. *On the Origin of Objects*. MIT Press, 1996.

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

[40] P. Cattin, D. Zlatnik, and R. Borer. *Biometric System using Human Gait*. Mechatronics and Machine Vision in Practice (M2VIP), 2001.

[41] P. Cavanagh. The language of vision. In *Proc. European Conference on Visual Perception*, 2003.

[42] C. Chang. The state of the art in video understanding. In *Proc. Int. Conference on Video Detection and Recognition*, 2001.

[43] N. Chang and K. Fu. Picture query languages for pictorial database systems. *IEEE Computer*, 14, 11:23–33, 1981.

[44] S. Chang, W. Chen, and H. Sundaram. Semantic visual templates: Linking visual features to semantics. In *Proc. Workshop on Content Based Video Search and Retrieval*, 1998.

[45] S.-F. Chang, Q. Chen, H. Meng, H. Sundaram, and D. Zhong. VideoQ: An automated content based video search system using visual cues. In *ACM Multimedia*, pages 313–324, 1997.

[46] Y. Chen, Y. Rui, and T. Huang. JPDAF based HMM for real-time contour tracking. In *Proc. Conference on Computer Vision and Pattern Recognition*, 2001.

[47] T. Choudhury, J. Rehg, V. Pavlovic, and A. Pentland. Boosting and structure learning in dynamic Bayesian networks for audio-visual speaker detection. In *Proc. Int. Conference on Pattern Recognition*, 2002.

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

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

[50] R. Cole. *The Management and Visualisation of Document Collections Using Formal Concept Analysis*. PhD thesis, Griffith University, 2000.

[51] G. Cooper and E. Herskovits. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9:309–347, 1992.

[52] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. In *Proc. European Conference on Computer Vision*, 1998.

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

[54] B. Coyne and R. Sproat. WordsEye: an automatic text-to-scene conversion system. In *Proc. Int. Conference on Computer Graphics and Interactive Techniques (ACM SIGGRAPH)*, 2001.

[55] J. Crowley, J. Coutaz, and F. Berard. Things that see: Machine perception for human computer interaction. *Communications of the ACM*, 43(3):54–64, 2000.

[56] J. Crowley, J. Coutaz, G. Rey, and P. Reignier. Perceptual components for context aware computing. In *Proc. Ubicomp 2002*, 2002.

[57] J. Cutting and L. Kozlowski. Recognising friends by their walk: Gait perception without familiarity cues. *Bulletin of the Psychonomic Society*, 9, 1977.

[58] J. Cutting and L. Kozlowski. Recognising the sex of a walker from dynamic point-light displays. *Perception of psychophysics*, 21, 1977.

[59] T. Darrell, G. Gordon, M. Harville, and J. Woodfill. Integrated person tracking using stereo, color, and pattern detection. In *Proc. Conference on Computer Vision and Pattern Recognition*, 1998.

[60] H. Daschiel. *Advanced Methods for Image Information Mining System: Evaluation and Enhancement of User Relevance*. PhD thesis, Technical University of Berlin, 2004.

[61] F. De la Torre and M. Black. Robust principal component analysis for computer vision. In *Proc. International Conference on Computer Vision*, 2001.

[62] F. De la Torre and M. Black. Robust parameterized component analysis: Theory and applications to 2d facial appearance models. *Computer Vision and Image Understanding*, 2003.

[63] D. Dennett. *Minds, machines, and evolution*, chapter Cognitive Wheels: The Frame Problem of AI, pages 129–151. Cambridge University Press, 1984.

[64] J. Denzler, M. Zobel, and J. Triesch. Probabilistic integration of cues from multiple cameras. In *Proc. 4th Workshop on Dynamic Perception*, 2002.

[65] A. Dey. Understanding and using context. *Personal and Ubiquitous Computing*, 5(1):4–7, 2001.

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

[67] W. Dolan, S. Richardson, and L. Vanderwende. MindNet: acquiring and structuring semantic information from text. Technical Report MSR-TR-98-23, Microsoft Research, 1998.

[68] B. Draper, U. Ahlrichs, and D. Paulus. Adapting object recognition across domains: A demonstration. *Lecture Notes in Computer Science*, 2095:256–270, 2001.

[69] P. Duffett-Smith and G. Woan. GSM CURSOR. International Patent Specification No. 9519087.2.

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

[71] A. Ekin, A. Tekalp, and R. Mehrotra. Semantic video querying using an integrated semantic-syntactic model. In *Proc. International Conference on Image Processing*, 2002.

[72] A. Ekin, A.M. Tekalp, and R. Mehrotra. Extraction of semantic description of events using Bayesian networks. In *Proc. ICASSP*, 2001.

[73] P. Eklund and R. Cole. Structured ontology and information retrieval for email search and discovery. In *Proc. Int. Symposium on Foundations of Intelligent System*, 2002.

[74] T. Ellis. Performance evaluation of tracking and surveillance. In *IEEE Workshop on Performance Evaluation in Tracking and Surveillance*, 2002.

[75] T. Ellis and M. Xu. Object detection and tracking in an open and dynamic world. In *IEEE Workshop on Performance Evaluation in Tracking and Surveillance*, 2001.

[76] T. Erickson. Some problems with the notion of context-aware computing. *Communications of the ACM*, 45(2):102–104, 2002.

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

[78] J. Feldman. Regularity-based perceptual grouping. *Computational Intelligence*, 13(4):582–623, 1997.

[79] D. Fensel. *Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce.* Springer-Verlag New York, Inc., 2003.

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

[81] C. Fillmore, C. Baker, and H. Sato. The FrameNet database and software tools. In *Proc. Int. Conference on Language Resources and Evaluation*, 2002.

[82] P. Flach. On the state of the art in machine learning: A personal review. *Artificial Intelligence*, 13(1/2):199–222, 2001.

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

[84] N. Friedman and D. Koller. Being Bayesian about network structure. In *Proc. Conference on Uncertainty in Artificial Intelligence*, 2000.

[85] N. Friedman, K. Murphy, and S. Russell. Learning the structure of dynamic probabilistic networks. In *Proc. Conference on Uncertainty in Artificial Intelligence*, pages 139–147, 1998.

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

[87] C. Garcia and G. Tziritas. Face detection using quantized skin color regions merging and wavelet packet analysis. *IEEE Transactions on Multimedia*, 1(3):264–277, 1999.

[88] D. Gavrila. The visual analysis of human movement: A survey. *Computer Vision and Image Understanding: CVIU*, 73(1):82–98, 1999.

[89] Z. Geradts. *Content-Based Information Retrieval from Forensic Databases.* PhD thesis, University of Utrecht, 2002.

[90] I. Getting. The Global Positioning System. *IEEE Spectrum*, 30(12):36–47, 1993.

[91] Z. Ghahramani and M. Beal. Propagation algorithms for variational Bayesian learning. In *Proc. Neural Information Processing Systems*, pages 507–513, 2000.

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

[93] S. Gong, J. Ng, and J. Sherrah. On the semantics of visual behaviour, structured events and trajectories of human action. *Image and Vision Computing*, 20(12):873–888, 2002.

[94] S. Gong, A. Psarrou, I. Katsouli, and P. Palavouzis. Tracking and Recognition of Face Sequences. In *European Workshop on Combined Real and Synthetic Image Processing for Broadcast and Video Production*, 1994.

[95] A. Graves and S. Gong. Spotting scene change for indexing surveillance video. In *Proc. British Machine Vision Conference*, 2003.

[96] R. Gregory. *Eye and Brain.* Princeton University Press, fifth edition, 1997.

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

[98] N. Guarino, C. Masolo, and G. Vetere. Ontoseek: Content-based access to the web. *IEEE Intelligent Systems*, 14(3):70–80, 1999.

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

[100] A. Gmez-Prez and D. Manzano-Macho. A survey of ontology learning methods and techniques. Deliverable 1.5, OntoWeb Project, IST-2000-29243, 2003.

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

[102] A. Hanbury. Circular statistics applied to colour images. *8th Computer Vision Winter Workshop*, 2003.

[103] R. Harle and A. Hopper. Dynamic world models from ray-tracing. In *Proc. Int. Conference on Pervasive Computing and Communications*, 2004.

[104] S. Harnad. The symbol grounding problem. *Physica D*, 42:335–346, 1990.

[105] A. Harter, A. Hopper, P. Steggles, A. Ward, and P. Webster. The anatomy of a context-aware application. In *Mobile Computing and Networking*, pages 59–68, 1999.

[106] P. Hayes. On semantic nets, frames and associations. In *Proc. Int. Joint Conference on Artificial Intelligence*, pages 99–107, 1977.

[107] D. Heckerman. A tutorial on learning with Bayesian networks. Technical Report MSR-TR-95-06, Microsoft Research, 1996.

[108] D. Heckerman. A tutorial on learning with Bayesian networks. In M. Jordan, editor, *Learning in Graphical Models*. MIT Press, 1998.

[109] D. Heesch and S. Rueger. Performance boosting with three mouse clicks - relevance feedback for CBIR. In *Proc. European Conference on Information Retrieval Research*, 2003.

[110] J. Heikkila and O. Silven. A four-step camera calibration procedure with implicit image correction. In *Proc. International Conference on Computer Vision*, 1997.

[111] B. Heisele, T. Serre, M. Pontil, T. Vetter, and T. Poggio. Categorization by learning and combining object parts. In *Proc. Neural Information Processing Systems*, 2001.

[112] G. Hendrix, E. Sacerdoti, D. Sagalowicz, and J. Slocum. Developing a natural language interface to complex data. *ACM Trans. on Database Systems*, 3:105–147, 1978.

[113] G. Herzog and K. Rohr. Integrating vision and language: Towards automatic description of human movements. In *KI-95: Advances in Artificial Intelligence*, pages 257–268. Springer, 1995.

[114] G. Herzog and P. Wazinski. VIsual TRAnslator: Linking perceptions and natural language descriptions. *Artificial Intelligence Review*, 8(2-3):175–187, 1994.

[115] J. Hightower and G. Borriello. Location systems for ubiquitous computing. *IEEE Computer*, 34(8):57–66, 2001.

[116] J. Hightower, B. Brumitt, and G. Borriello. The location stack: A layered model for location in ubiquitous computing. In *Workshop on Mobile Computing Systems and Applications*, 2002.

[117] A. Hopper. Sentient Computing - The Royal Society Clifford Paterson Lecture. *Philosophical Transactions of the Royal Society of London*, 358(1773):2349–2358, 2000.

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

[119] M. Hu. Visual pattern recognition by moment invariants. *IRA Transactions on Information Theory*, 17(2):179–187, 1962.

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

[121] B. Huet and E. Hancock. Fuzzy relational distance for large-scale object recognition. In *Proc. Conference on Computer Vision and Pattern Recognition*, 1998.

[122] A. Hunter and L. Marten. Reasoning with structured text using world knowledge. Technical report, Department of Computer Science, University College London, 2001.

[123] S. Intille and A. Bobick. Representation and visual recognition of complex, multi-agent actions using belief networks. In *IEEE Workshop on the Interpretation of Visual Motion*, 1998.

[124] S. Ioffe and D. Forsyth. Human tracking with mixtures of trees. In *Proc. International Conference on Computer Vision*, 2001.

[125] S. Ioffe and D. Forsyth. Mixtures of trees for object recognition. In *Proc. Conference on Computer Vision and Pattern Recognition*, 2001.

[126] D. Ipina. *Visual Sensing and Middleware Support for Sentient Computing.* PhD thesis, Cambridge University Engineering Department, 2002.

[127] D. Ipina and A. Hopper. TRIP: a low-cost vision-based location system for ubiquitous computing. *Personal and Ubiquitous Computing*, 6(3):206–219, 2002.

[128] M. Isard and A Blake. Condensation – conditional density propagation for visual tracking. *International Journal of Computer Vision*, 29(1):5–28, 1998.

[129] M. Isard and A. Blake. ICONDENSATION: Unifying low-level and high-level tracking in a stochastic framework. *Lecture Notes in Computer Science*, 1406, 1998.

[130] Y. Ivanov and A. Bobick. Recognition of visual activities and interactions by stochastic parsing. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(8), 2000.

[131] J. J. Vermaak, P. Perez, M. Gangnet, and A. Blake. Towards improved observation models for visual tracking: selective adaptation. In *Proc. European Conference on Computer Vision*, 2002.

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

[133] A. Jaimes and S. Chang. A conceptual framework for indexing visual information at multiple levels. In *IS&T SPIE Internet Imaging*, 2000.

[134] A. Jaimes and S.-F. Chang. Integrating multiple classifiers in visual object detectors learned from user input. In *Proc. Asian Conference on Computer Vision*, 2000.

[135] F.V. Jensen. *An Introduction to Bayesian Networks.* Springer Verlag, 1996.

[136] A Jepson, D. Fleet, and T. El-Maraghi. Robust online appearance models for visual tracking. In *Proc. Conference on Computer Vision and Pattern Recognition*, 2001.

[137] I. Jermyn, D. Jacobs, and D. Geiger. Target-driven perceptual grouping. In *Proc. IEEE Workshop on Perceptual Organization in Computer Vision*, 1998.

[138] M. Johnston. Multimodal language processing. In *Proc. Int. Conference on Spoken Language Processing*, 1998.

[139] M. Johnston and S. Bangalore. Finite-state multimodal parsing and understanding. In *Proc. of COLING'2000.* Association for Computational Linguistics, 2000.

[140] N. Jojic and B. Frey. Learning flexible sprites in video layers. In *Proc. Conference on Computer Vision and Pattern Recognition*, 2001.

[141] M. Jordan, editor. *Learning in Graphical Models.* MIT Press, 1999.

[142] S. Ju, M. Black, S. Minneman, and D. Kimber. Summarization of videotaped presentations: Automatic analysis of motion and gesture. *IEEE Trans. on Circuits and Systems for Video Technology*, 8(5):686–696, 1998.

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

[144] B. Katz, J. Lin, C. Stauffer, and E. Grimson. Answering questions about moving objects in surveillance videos. In *Proc of the AAAI Spring Symposium on New Directions in Question Answering*, 2003.

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

[146] Y. Keselman and S. Dickinson. Bridging the representation gap between models and exemplars. In *IEEE Workshop on Models versus Exemplars in Computer Vision*, 2001.

[147] K. Koffka. *Principles of Gestalt Psychology.* Harcourt Brace, New York, 1935.

[148] A. Kojima, T. Tamura, and K. Fukunaga. Natural language description of human activities from video images based on concept hierarchy of actions. *Int. Journal of Computer Vision*, 50(2):171–184, 2002.

[149] M. Kokar and J. Wang. An example of using ontologies and symbolic information in automatic target recognition. In *Proc. SPIE Sensor Fusion: Architectures, Algorithms, and Applications VI*, pages 40–50, 2002.

[150] H. Kollnig, H.-H. Nagel, and M. Otte. Association of motion verbs with vehicle movements extracted from dense optical flow fields. In *Proc. European Conference on Computer Vision*, 1994.

[151] U. Kruschwitz. Exploiting structure for intelligent web search. In *Proc. Int. Conference on System Sciences*, Maui, Hawaii, 2001.

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

[153] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proc of the IEEE*, 86(11):2278–2324, 1998.

[154] J. Lee and R. Scholtz. Ranging in a Dense Multipath Environment Using an UWB Radio Link. *IEEE Journal on Selected Areas in Communications*, 20(9), 2002.

[155] M. Lee and I. Cohen. Human upper body pose estimation in static images. In *Proc. European Conference on Computer Vision*, 2004.

[156] D. Lenat. CYC: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11), 1995.

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

[158] R. Lienhart, A. Kuranov, and V. Pisarevsky. Empirical analysis of detection cascades of boosted classifiers for rapid object detection. Technical report, MRL, Intel Corporation, Santa Clara, CA 95052, USA, 2002.

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

[160] X. Liu and T. Chen. Video-based face recognition using adaptive Hidden Markov models. In *Proc. Conference on Computer Vision and Pattern Recognition*, 2003.

[161] S. Loncaric. A survey of shape analysis techniques. *Pattern Recognition*, 31(8):983–1001, 1998.

[162] J. Lowe, C. Baker, and C. Fillmore. A frame-semantic approach to semantic annotation. In *SIGLEX workshop on Tagging Text with Lexical Semantics*, 1997.

[163] Yang M., D. Kriegman, and Ahuja N. Detecting faces in images: A survey. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(1):34–58, 2002.

[164] D. MacKay. Ensemble learning and evidence maximization. In *Proc. Neural Information Processing Systems*, 1995.

[165] D. Magee. A qualitative, multi-scale grammar for image description and analysis. In *Proc. British Machine Vision Conference*, 2002.

[166] R. Mann, A. Jepson, and J. Siskind. The computational perception of scene dynamics. *Computer Vision and Image Understanding*, 65(2):113–128, 1997.

[167] D. Marr. *Vision*. Freeman, San Francisco, 1982.

[168] P. Martin. *Web Intelligence*, chapter Knowledge Representation, Sharing and Retrieval on the Web, pages 243–276. Springer-Verlag, 2003.

[169] S. Maskell, M. Rollason, N. Gordon, and D. Salmond. Efficient particle filtering for multiple target tracking with application to tracking in structured images. *Image and Vision Computing*, 21:931–939, 2003.

[170] J. McCarthy and P. Hayes. Some philosophical problems from the standpoint of artificial intelligence. *Machine Intelligence*, 4, 1969.

[171] S. McKenna, Y. Raja, and S. Gong. Object tracking using adaptive color mixture models. In *Proc. Asian Conference on Computer Vision*, pages 615–622, 1998.

[172] M. Mechkour, C. Berrut, and Y. Chiaramella. Using conceptual graph framework for image retrieval. In *Proc. Int. Conference on Multi-Media Modeling*, 1995.

[173] A. Menezes. Better contextual translation using machine learning. In *Proc. 5th Conference of the Association for Machine Translation in the Americas*, 2002.

[174] V. Mezaris, I. Kompatsiaris, and M. Strintzis. An ontology approach to object-based image retrieval. In *Proc. International Conference on Image Processing*, 2003.

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

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

[177] T. Moeslund and E. Granum. A survey of computer vision-based human motion capture. *Computer Vision and Image Understanding: CVIU*, 81(3):231–268, 2001.

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

[179] D. Moore and I. Essa. Recognizing multitasked activities using stochastic context-free grammar. In *Proc. Workshop on Models vs Exemplars in Computer Vision*, 2001.

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

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

[182] K. Murphy. The Bayes Net Toolbox for Matlab. *Computing Science and Statistics*, 33, 2001.

[183] K. Murphy. *Dynamic Bayesian Networks: Representation, Inference and Learning*. PhD thesis, University of California at Berkeley, 2002.

[184] H.-H. Nagel. From image sequences towards conceptual descriptions. *Image and Vision Computing*, 6:59–74, 1988.

[185] H.-H. Nagel. A vision of 'vision and language comprises action': An example from road traffic. *Artificial Intelligence Review*, 8:189–214, 1994.

[186] S. Nepal, M. Ramakrishna, and J. Thom. A fuzzy object query language (FOQL) for image databases. In *Proc. Int. Conference on Database Systems for Advanced Applications*, 1999.

[187] R. Nevatia, J. Hobbs, and B. Bolles. An ontology for video event representation. In *Proc. Int. Workshop on Detection and Recognition of Events in Video (at CVPR04)*, 2004.

[188] R. Nevatia, T. Zhao, and S. Hongeng. Hierarchical language-based representation of events in video streams. In *Proc. IEEE Workshop on Event Mining*, 2003.

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

[190] J. Nord, K. Synnes, and P. Parnes. An architecture for location aware applications. In *Proc of the Hawaii Int. Conference on System Sciences*, 2002.

[191] K. Nummiaro, E. Koller-Meier, and L.V. Gool. An adaptive color-based particle filter. *Image and Vision Computing*, 21:99–110, 2003.

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

[193] K. Okuma, A. Taleghani, N. De Freitas, J. Little, and D. Lowe. A boosted particle filter: Multitarget detection and tracking. In *Proc. European Conference on Computer Vision*, 2004.

[194] N. Oliver, B. Rosario, and A. Pentland. A Bayesian computer vision system for modeling human interactions. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(8):831–843, 2000.

[195] B. Omelayenko. Machine learning for ontology learning. Presentation at the Int. Jyvaskyla Summer School, Finland, 2000.

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

[197] K. Pastra, H. Saggion, and Y. Wilks. Extracting relational facts for indexing and retrieval of crime-scene photographs. *IEEE Intelligent Systems*, 18(1):55–61, 2002.

[198] D. Paulus, U. Ahlrichs, B. Heigl, J. Denzler, J. Hornegger, and H. Niemann. Active knowledge-based scene analysis. In *Proc. Int. Conference on Vision Systems*, pages 180–199, 1999.

[199] V. Pavlovic, A. Garg, J. Rehg, and T. Huang. Multimodal speaker detection using error feedback dynamic Bayesian networks. In *Proc. Conference on Computer Vision and Pattern Recognition*, 2000.

[200] V. Pavlovic, J. Rehg, T.-J. Cham, and K. Murphy. A dynamic Bayesian network approach to figure tracking using learned dynamic models. In *Proc. International Conference on Computer Vision*, 1999.

[201] J. Peng and B. Bhanu. Closed-loop object recognition using reinforcement learning. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 20(2), 1998.

[202] P. Perez, J. Vermaak, and A. Blake. Data fusion for visual tracking with particles. In *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2004.

[203] J. Peterson, K. Mahesh, and A. Goel. Situating natural language understanding within experience-based design. Technical Report GIT-CC-94-18, College of Computing, Georgia Institute of Technology, 1994.

[204] E. Petrakis and C. Faloutsos. Similarity searching in large image databases. Technical Report CS-TR-3388, Department of Computer Science, College Park, MD, USA, 1994.

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

[206] A. Pfeffer and D. Koller. Semantics and inference for recursive probability models. In *Proc. AAAI'00*, 2000.

[207] A. Pfeffer, D. Koller, B. Milch, and K. Takusagawa. SPOOK: A system for probabilistic object-oriented knowledge representation. In *Proc. Conference on Uncertainty in AI*, 1999.

[208] J. Piater and J. Crowley. Multi-modal tracking of interacting targets using gaussian approximations. In *IEEE Workshop on Performance Evaluation in Tracking and Surveillance*, 2001.

[209] C. Pinhanez and A. Bobick. Approximate world models: Incorporating qualitative and linguistic information into vision systems. In *AAAI'96*, 1996.

[210] K. Popper. *Conjectures and Refutations: The Growth of Scientific Knowledge.* London: Routledge & Kegan Paul, 1963.

[211] A. Prati, I. Mikic, R. Cucchiara, and M. Trivedi. Comparative evaluation of moving shadow detection algorithms. In *Proc. Workshop on Empirical Evaluation Methods in Computer Vision*, 2001.

[212] N. Priyantha, K. Allen, H. Balakrishnan, and S. J. Teller. The cricket compass for context-aware mobile applications. In *Mobile Computing and Networking*, pages 1–14, 2001.

[213] L. Quan and Z.-D. Lan. Linear N-point camera pose determination. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 21(8):774–780, 1999.

[214] W. Quine. *From a Logical Point of View*, chapter On What There Is. Harper and Row, New York, 1953.

[215] L. Rabiner. A tutorial on Hidden Markov models and selected applications in speech recognition. *Proc of the IEEE*, 77(2), 1989.

[216] D. Ramanan and D. Forsyth. Finding and tracking people from the bottom up. In *Proc. Conference on Computer Vision and Pattern Recognition*, 2003.

[217] J. Rehg, K. Murphy, and P. Fieguth. Vision-based speaker detection using Bayesian networks. In *Proc. Conference on Computer Vision and Pattern Recognition*, 1999.

[218] P. Remagnino, T. Tan, and K. Baker. Agent orientated annotation in model based visual surveillance. In *Proc. International Conference on Computer Vision*, 1998.

[219] R. Rimey. *Control of Selective Perception using Bayes Nets and Decision Theory.* PhD thesis, University of Rochester Computer Science Department, 1993.

[220] S. Robertson and K. Sparck-Jones. Simple, proven approaches to text retrieval. Technical Report 356, Cambridge University Computer Laboratory, 1997.

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

[222] K. Rodden. *Evaluating similarity-based visualisations as interfaces for image browsing.* PhD thesis, Cambridge University Computer Laboratory, 2001.

[223] C. Rosenberg and M. Hebert. Training object detection models with weakly labeled data. In *Proc. British Machine Vision Conference*, 2002.

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

[225] N. Rowe and B. Frew. *Automatic classification of objects in captioned descriptive photographs for retrieval*, chapter 4, pages 65–79. AAAI Press, 1997.

[226] S. Roweis and Z. Ghahramani. A unifying review of linear Gaussian models. *Neural Computation*, 11(2):305–345, 1999.

[227] D. Roy. Learning visually grounded words and syntax of natural spoken language. *Evolution of Communication*, 4, 2001.

[228] D. Roy. A trainable visually-grounded spoken language generation system. In *Proc. Int. Conference of Spoken Language Processing*, 2002.

[229] Y. Rubner, C. Tomasi, and L. Guibas. A metric for distributions with applications to image databases. In *Proc. International Conference on Computer Vision*, 1998.

[230] W. Rucklidge. Locating objects using the Hausdorff distance. In *Proc. International Conference on Computer Vision*, 1995.

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

[232] S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, second edition, 2002.

[233] J. Ryu and S.-B. Cho. Gender recognition of human behaviors using neural ensembles. In *Proc. INNS/IEEE Int. Joint Conference on Neural Networks*, 2001.

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

[235] S. Sarkar and K.L. Boyer. Perceptual organization in computer vision: A review and a proposal for a classificatory structure. *SMC*, 23:382–399, 1993.

[236] R. Schapire. Theoretical views of boosting and applications. In *Proc. Int. Conference on Algorithmic Learning Theory*, 1999.

[237] B. Schilit, N. Adams, and R. Want. Context-Aware Computing Applications. In *Proc. Workshop on Mobile Computing Systems and Applications*, 1994.

[238] A. Senior, A. Hampapur, Y-L. Tian, L. Brown, S. Pankanti, and R. Bolle. Appearance models for occlusion handling. In *IEEE Workshop on Performance Evaluation in Tracking and Surveillance*, 2001.

[239] L. Shapiro and G. Stockman. *Computer Vision*. Prentice Hall, 2001.

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

[241] J. Sherrah and S. Gong. Tracking discontinuous motion using Bayesian inference. In *Proc. European Conference on Computer Vision*, pages 150–166, 2000.

[242] J. Sherrah and S. Gong. Continuous global evidence-based Bayesian modality fusion for simultaneous tracking of multiple objects. In *Proc. International Conference on Computer Vision*, 2001.

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

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

[245] D. Skocaj and A. Leonardis. Robust continuous subspace learning and recognition. In *Proc. Int. Electrotechnical and Computer Science Conference*, 2002.

[246] A. Smailagic, D. Siewiorek, and J. Anhalt. Towards context aware computing: Experiences and lessons learned. *IEEE Journal on Intelligent Systems*, 16(3):38–46, 2001.

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

[248] B. Smith. *Formal Ontology and Information Systems*, chapter Ontology and Information Systems. ACM Press, 2001.

[249] B. Smith and C. Welty. Ontology—towards a new synthesis. In *Proc. Int. Conference on Formal Ontology in Information Systems*, 2001.

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

[251] P. Smith. *Edge-based Motion Segmentation.* PhD thesis, Cambridge University Engineering Department, 2001.

[252] G. Socher, G. Sagerer, and P. Perona. Bayesian reasoning on qualitative descriptions from images and speech. *Image and Vision Computing*, 18(2):155–172, 2000.

[253] J. Sowa. *Knowledge Representation: Logical, Philosophical, and Computational Foundations*. Brooks Cole Publishing, 1999.

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

[255] K. Spark-Jones and P. (editors) Willett. *Readings in Information Retrieval*. Morgan Kaufmann Publishers, 1997.

[256] M. Spengler and B. Schiele. Towards robust multi-cue integration for visual tracking. *Lecture Notes in Computer Science*, 2095:93–106, 2001.

[257] R. Srihari. Computational models for integrating linguistic and visual information: A survey. *Artificial Intelligence Review, special issue on Integrating Language and Vision*, 8:349–369, 1995.

[258] S. Staab and R. Studer, editors. *Handbook on Ontologies*. International Handbooks on Information Systems. Springer, 2004.

[259] F. Stajano. *Security for Ubiquitous Computing*. John Wiley and Sons, 2002.

[260] C. Stauffer and E. Grimson. Learning patterns of activity using real-time tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):747–757, 2000.

[261] S. Stillman and I. Essa. Towards reliable multimodal sensing in aware environments. In *Proc. Perceptual User Interfaces Workshop, ACM UIST 2001*, 2001.

[262] B. Stroustrup. What is "object-oriented programming"? In *Proc. 1st European Software Festival*, 1991.

[263] L. Tang, R. Hanka, H. Ip, K. Cheung, and R. Lam. Semantic query processing and annotation generation for content-based retrieval of histological images. In *Proc. SPIE Medical Imaging*, 2000.

[264] M. Thonnat and N. Rota. Image understanding for visual surveillance applications. In *Proc. of 3rd Int. Workshop on Cooperative Distributed Vision*, 1999.

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

[266] M. Tipping. The relevance vector machine. *Advances in Neural Information Processing Systems*, 12, 2000.

[267] A. Torralba, K. Murphy, W. Freeman, and A. Mark. Context-based vision system for place and object recognition. In *Proc. International Conference on Computer Vision*, 2003.

[268] C.P. Town. Adaptive integration of visual tracking modalities for sentient computing. In *Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, 2003.

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

[270] C.P. Town and D.A. Sinclair. 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.

[271] K. Toyama and A. Blake. Probabilistic tracking in a metric space. In *Proc. International Conference on Computer Vision*, 2001.

[272] K. Toyama and E. Horvitz. Bayesian modality fusion: Probabilistic integration of multiple vision algorithms for head tracking. In *Proc. Asian Conference on Computer Vision*, 2000.

[273] K. Toyama and Y. Wu. Bootstrap initialization of nonparametric texture models for tracking. In *Proc. European Conference on Computer Vision*, pages 119–133, 2000.

[274] J. Triesch and C. von der Malsburg. Self-organized integration of visual cues for face tracking. In *Proc. Int. Conference on Automatic Face and Gesture Recognition*, pages 102–107, 2000.

[275] B. Triggs, A. Zisserman, and R. Szeliski, editors. *Vision Algorithms: Theory and Practice*. Number 1883 in Lecture Notes in Computer Science. Springer, 1999.

[276] W. Tsai and K. Fu. Attributed grammars - a tool for combining syntactic and statistical approaches to pattern recognition. *IEEE Transactions on Systems, Man and Cybernetics*, SMC-10(12), 1980.

[277] J. Tsotsos, J. Mylopoulos, H. Covvey, and S. Zucker. 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.

[278] M. Uschold and M. Gruninger. Ontologies: principles, methods, and applications. *Knowledge Engineering Review*, 11:93–155, 1996.

[279] A. Vailaya, H. Zhang, C. Yang, F. Liu, and A. Jain. Automatic image orientation detection. *IEEE Transactions on Image Processing*, 11(7):746–755, 2002.

[280] V. Vapnik. *Statistical learning theory*. John Wiley & Sons, 1998.

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

[282] P. Viola and M. Jones. Fast and robust classification using asymmetric Adaboost and a detector cascade. In *Proc. Neural Information Processing Systems*, 2001.

[283] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proc. International Conference on Computer Vision*, 2001.

[284] J. Vogel and B. Schiele. On performance categorization and optimization for image retrieval. In *Proc. European Conference on Computer Vision*, 2002.

[285] S. Wachsmuth, G. Socher, H. Brandt-Pook, F. Kummert, and G. Sagerer. Integration of vision and speech understanding using Bayesian networks. *Videre: A Journal of Computer Vision Research*, 1(4), 2000.

[286] W. Wahlster. One word says more than a thousand pictures - on the automatic verbalization of the results of image sequence analysis systems. *Computers and Artificial Intelligence*, 8:479–492, 1989.

[287] Y. Wang and H. Zhang. Detecting image orientation based on low-level visual content. *Computer Vision and Image Understanding*, 93(3), 2004.

[288] R. Want, K. Fishkin, A. Gujar, and B. L. Harrison. Bridging Physical and Virtual Worlds with Electronic Tags. In *Proc. CHI'99*, 1999.

[289] R. Want, A. Hopper, V. Falcao, and J. Gibbons. The Active Badge location system. Technical Report 92.1, AT&T Laboratories Cambridge, 1992.

[290] M. Weiser. The Computer for the 21st Century. *Scientific American*, 1991.

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

[292] M. Wertheimer. Principles of perceptual organisation. In W.H. Ellis, editor, *Source Book of Gestalt Psychology.* Harcourt Brace, New York, 1938.

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

[294] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland. Pfinder: Real-time tracking of the human body. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(7):780–785, 1997.

[295] Y. Wu and T. Huang. Vision-based gesture recognition: A review. *Lecture Notes in Computer Science*, 1739:103–114, 1999.

[296] Y. Wu and T. Huang. A co-inference approach to robust visual tracking. In *Proc. International Conference on Computer Vision*, 2001.

[297] C. Xu and J. Prince. Snakes, shapes, and gradient vector flow. *IEEE Transactions on Image Processing*, 7(3):359–369, 1998.

[298] Z. Zhang. Flexible camera calibration by viewing a plane from unknown orientations. In *Proc. International Conference on Computer Vision*, 1999.

[299] R. Zhao and W. Grosky. From features to semantics: Some preliminary results. In *Proc. IEEE Int. Conference on Multimedia and Expo*, pages 679–682, 2000.

[300] S. Zhou, V. Krueger, and R. Chellappa. Probabilistic recognition of human faces from video. In *Computer Vision and Image Understanding*, 2003.

[301] D. Zotkin, R. Duraiswami, H. Nanda, and L. Davis. Multimodal tracking for smart videoconferencing. In *Proc. 2nd Int. Conference on Multimedia and Expo*, 2001.