Advanced Visual Surveillance using Bayesian Networks

Hilary Buxton

School of Cognitive and Computing Sciences
University of Sussex
Falmer, Brighton BN1 9QH, UK

Shaogang Gong

Department of Computer Science
QMW, University of London
Mile End Road, London E1 4NS, UK

Abstract

Advanced visual surveillance systems not only need to track moving objects but also interpret their patterns of behaviour. This means that solving the information integration problem becomes very important. We use conceptual knowledge of both the scene and the visual task to provide constraints. We also control the system using dynamic attention and selective processing. Bayesian belief network (BBN) techniques support this as well as allowing us to model dynamic dependencies between parameters involved in visual interpretation. We illustrate these arguments using experimental results from a traffic surveillance application. In particular, we show that using expectations of object trajectory, size and speed for the particular scene can improve robustness and sensitivity in dynamic tracking and segmentation. We also show that behavioural evaluation under attentional control can be achieved using a combination of a static BBN tasknet and dynamic network (DBN). The causal structure of these networks provides a framework for the design and integration of advanced vision systems.

1 Introduction

Visual surveillance primarily involves the interpretation of image sequences. Advanced visual surveillance goes further and automates the detection of predefined alarm events in a given context. However, it is the intelligent, dynamic scene and event discrimination which lies at the heart of advanced vision systems. Developing a systematic methodology for the design, implementation and integration of such systems is currently a very important research problem [3, 5, 14, 30, 40]. These methods must take into account the fact that vision is a computationally difficult problem as information available in the image does not provide a one-to-one mapping to physical objects in space. In fact, visual evidence extracted by machine-based processing is almost always subject to uncertainty and incompleteness due to noise, occlusion, and the general ill-posed nature of the inverse-perspective projection used to infer the scene from the image data. One way of overcoming some of these problems is to build in more knowledge of the scene and tasks so the primary intention of our method is to allow the representation of conceptual knowledge in a readily accessible form at all levels of visual processing. For example, in object detection and tracking in the image, we have demonstrated that it is beneficial to bring scene-based knowledge of expected object trajectories, size and speed into the interpretation process [17, 18, 19]. We have also shown that both scene and task-based knowledge allows for selective processing under attentional control for behavioural evaluation [20, 21, 22]. However, it remains to achieve all this in a tightly coupled scheme that allows for greater computational efficiency in performing multiple visual tasks.

In addition to this general requirement for integration of information in advanced visual surveillance, we have adopted more specific requirements. A fixed, precalibrated camera model and precomputed ground-plane geometry is used to simplify the interpretation of the scene data in the on-line system. We also adopt a knowledge-based approach in which domain specific models of the dynamic objects, events and behaviour are used to meet the requirement for sensitive and accurate performance. The requirements for the camera model are to support accurate mappings from 2D to 3D and vice-versa, while those for the ground-plane representation are inherited from the kind of behavioural analysis we need to support. We require both metrical and topological information to be recovered from our ground-plane representation as well as needing access to semantic and structural properties that determine the behaviour of the dynamic objects. The representation and reasoning for the dynamic objects needs 3D position, orientation and occupancy in tracking as well as recognition of object type. The requirements for the event and behaviour analysis on the other hand are less easily defined as they depend on the range of tasks to be supported. We adopt the assumption that a flexible set of behavioural evaluations and status reports may be requested which require compositional analysis in terms of the events observed. This decomposition, in which perceptual processing first recovers trajectory-based descriptions of the dynamic objects and is followed by conceptual processing, leads to a combinatorial explosion if purely data-driven control is used. Therefore, we return to the general requirement for attentional, purposive (or task-based) integration and control in advanced surveillance systems.

Computer vision, under the influence of Marr [27], has developed sophisticated algorithmic procedures

for individual visual competences. For a single visual task such as dynamic object recognition, it is possible to integrate such systems [29]. However, it is a clumsy approach to building advanced vision systems that are required to perform multiple tasks. In recent years, Ullman [41] has argued for the importance of integration of multiple visual routines. Our approach owes much to Ballard who proposed an animate vision approach [5] for two reasons: first, vision is better understood in the context of the visual behaviours engaging the system without requiring detailed internal representations of the scene; and second, it is important to have a system framework that integrates visual processing within the task context. These arguments are supported by Bajcsy and Allen's experiments in active vision [4] and the further demonstration by Aloimonos *et al.* [3] that some of the intrinsically ill-conditioned processes required for the reconstruction of physical features become stable when active vision is used. In our work, we have adapted active vision techniques to surveillance tasks where identifying "what","where" and "when" is just as essential for effective and efficient performance [19, 22].

In visual surveillance we often need knowledge-based techniques which are less generic but allow us to obtain robust and accurate results for a particular domain. The ACRONYM system of Brooks [7] used symbolic reasoning to analyse static scenes in a cycle of prediction, description, and interpretation. Such explicit reasoning about constraints in not able to deliver the fast performance required for dynamic visual surveillance. Recent work has improved the computational tractability by applying BBNs and decision theories [6, 25]. However, these studies do not address the specific computational difficulties involved in the interpretation of image sequences. Recent work using model-based approaches in image sequence analysis [24, 45] does effectively address the issue of dynamic interpretation. However, they do not compute behavioural descriptions or use task-dependent processing. We could use an approach based on database query techniques to deliver task-dependent behavioural descriptions. For example [10] proposed a scheme for incorporating explicit spatio-temporal knowledge into a Query-Based Vision System for understanding biological image sequences. However, this kind of system performs off-line processing and uses a fixed set of parameters for the objects of interest. Here, however, we want to interpret the spatio-temporal interactions between observed objects on-line while using the scene as its own best memory. To build such vision systems we need to address the question of how knowledge can be mapped onto computation to dynamically deliver consistent interpretations. This involves a more fundamental analysis of the spatio-temporal regularities in the image data so that we can exploit them as constraints in the processing scheme. We recently developed a Bayesian network approach [19] as it can support this kind of effective knowledge representation. It also provides the means for solving the information integration problem which is central to building robust systems capable of working on real image sequence data.

Pioneering work by Nagel [30] and Neumann [31] has emphasised the need to deliver conceptual or symbolic descriptions of behaviour from image sequences. More immediate background here, however, is the investigation of methods for real-time knowledge-based vision in the ESPRIT project VIEWS. System level integration of perceptual processing with conceptual understanding of traffic scenes allowed the development of three working demonstrators: airport stand area surveillance; multi-band tracker for airport ground traffic; and incident detection for road traffic scenes [12]. The conceptual processing in these systems was designed to handle missing information in the perceptual output as well as coping with behavioural variability for objects in the scenes. However, problems remain as only highly constrained feedback of information from the conceptual processing to the perceptual level was implemented in the VIEWS demonstrators. This was mainly concentrated on occlusion handling where it was vital for the system to consistently relabel emerging vehicles [37].

In what follows, we briefly review the system components and go on to present the basics of Bayesian nets and associated belief revision. Then we discuss our experimental results from focussed segmentation and tracking of moving objects and from selective task-based attentional control in behavioural evaluation. We conclude with a summary and suggestions for future work.

2 System Components

In the VIEWS project, the components required for an effective visual surveillance system have been identified. To simplify the run-time system we assume a precalibrated camera model, precomputed ground-plane map, as well as a set of object, event and behaviour models. We can then characterise the perceptual processing in such a system as providing track descriptions for the moving objects in the scene together with suggested object type using 2D image motion and 3D model-based vision techniques which are based on the camera, ground-plane, and object models. The conceptual processing can be characterised as taking these descriptions and then converting them into a consistent behavioural description using AI techniques based on the event and behaviour models as well as the camera and ground-plane models which define our field of view. In the following, we briefly examine five essential aspects of data representation and computation in our visual surveillance systems. These are: 1) camera models and their calibration where we emphasise design choices for off-line fixed camera surveillance; 2) ground-plane representation where we discuss requirements for supporting behavioural analysis using a cellular decomposition of space; 3) object recognition where we discuss the advantages of volumetric model representations for our application; 4) tracking dynamic objects where we discuss the need to fully integrate the motion analysis in model-based tracking schemes; and 5) behavioural representation and analysis where we again emphasise the need for appropriate techniques for on-line analysis and introduce cellular decomposition of time.

2.1 Camera Models and Calibration

The design choices for building systems are fundamentally guided by the requirements of the tasks we have to accomplish. In visual surveillance, two information transformations are essential: 1) infer 3D measurements from 2D image features through inverse-perspective projection and 2) predict the existence of 2D features for 3D object hypotheses. Such transformations are determined by calibrated camera parameters. However, decisions about the type of camera calibration must take into account the fact that we use a wide-angle lens to capture the activity over a wide-area scene. Camera calibration techniques have been established across a range of requirements for accuracy and efficiency [38, 42]. However, surveillance based on a wide-angle, static camera means that overcoming non-linear distortion is significant whilst dynamic calibration is less important. Since these operations are computed off-line, and it is only the resulting geometry that is used on-line, the modelling can be quite elaborate to allow accurate mappings from 2D to 3D and vice versa.

2.2 Ground-plane Representation

The representation of the ground-plane knowledge in our system needs to be very closely bound to the behavioural analysis we need to support for our surveillance system. Although there are a wide range of spatial representation and reasoning methods reported in the literature that have been developed to support a variety of different purposes, we need a representation that is closely tailored to our requirements. In surveillance, we require both metrical information such as angles and distances between spatial primitives and topological information such as neighbours and enclosure relationships. The behavioural analysis can be facilitated by regarding the problem as interpreting the motion patterns over time within a framework provided by a static environment. It also helps to have semantic as well as structural properties made explicit in our representation as these shape the behaviour of our purposively moving objects. For example, in the road traffic domain, we need to consider not only the lane boundaries and direction of traffic flow but also the "give-way" regions.

The spatial representation and reasoning in our surveillance system, then, must support: 1) the description of the static environment in the field of view, 2) the spatial occupancy of a moving object relative to this environment including its instantaneous position, extent and the region it occupies, 3) the spatial organisation of these objects at a given moment with respect to the environment and each other, and 4) an understanding of what the different regions in the environment "mean" in terms of physical and semantic constraints. The semantic constraints, such as possible paths through the environment, are represented as they are effective in the interpretation of observed behaviour in line with our purposive design strategy. We also need to consider that interpretations can operate either in the image-plane or on the ground-plane projection which provides an overhead view (figure 1). Both types of reasoning are essentially 2D and involve both metrical and topological relationships. However, reasoning on the ground-plane will require run-time 3D model-based reasoning together with the camera model to get the necessary ground-plane projections of object position, orientation and extent. We can then integrate knowledge of the ground-plane in the motion tracking using a precomputed projection into the image-plane to provide prior expectations of object trajectory, speed and size.

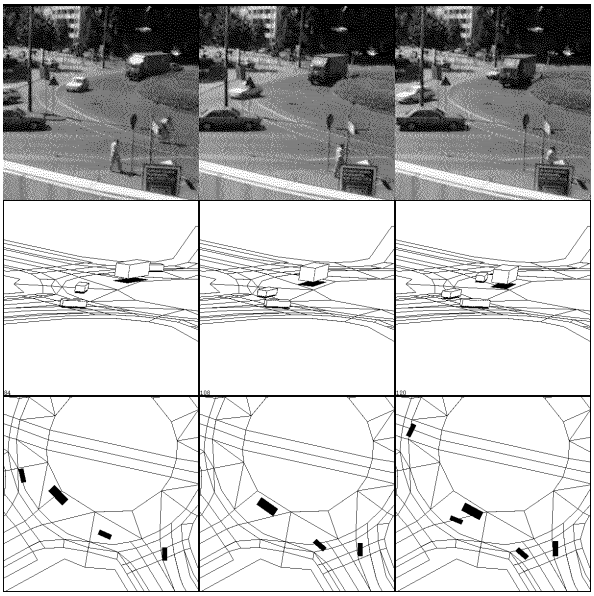

Figure 1: *The image plane, 3D space with dynamic objects, and the ground-plane representation of a traffic roundabout.*

Representations of space used in intermediate visual processing are concerned with supporting the immediate requirements of the task and include geometric and topological approaches. In surveillance we are concerned with perception of the moving object in the ground-plane context and need a wide-ranging model of space. An important development has been Fleck's cellular topology [15, 16] which can support the representation of digitised spaces for edge detection and stereo matching as well as applications in qualitative physics and spatial descriptions in natural language. This representation has been extended and tailored to the needs of surveillance by Howarth and Buxton [21]. The cellular decomposition underlying this approach supports both the metrical and topological properties we require in the ground-plane framework for interpreting the behaviour of moving objects. The cells can be made to conform to the ground-plane layout and grouped into a hierarchy of regions supporting the meaningful representation of spatial context for the on-line processing of behavioural descriptions (figure 2). The decomposition is obtained by using a map editor to intersect (1) the road surface into (2) the entry and exit roads, (3) the turn-right zone, (4) the roundabout, (5) the lanes of the entry road, (6) the

give-way zone, (7)the turning-zone, (8) the give-way-to zone, and (9) the leaf regions. The regular cells subdivide these regions to give a metrical position and support the intersection and extrusion processes used for ground-plane predictions. Spatially invariant behavioural information such as static give-way regions can then be used in the contextual indexing when we need to describe what the objects are doing in the scene. This extended cellular representation has been used to support full event and behaviour evaluation in both passive and attentionally controlled surveillance systems [21, 22].

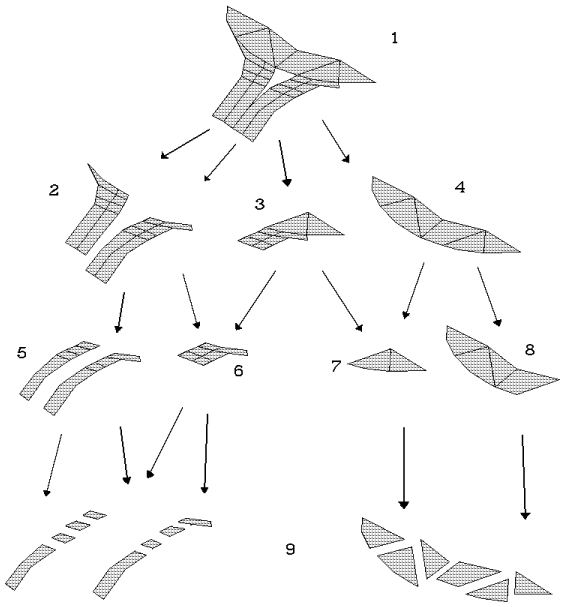

Figure 2: *Hierarchical decomposition into regions and cells representing a part of the roundabout ground-plane.*

2.3 Object Recognition

The overall purpose of a visual surveillance system, as we have emphasised, is to provide meaningful descriptions of purposively moving objects. This means coherent, effective and sufficient interpretations of dynamic behavioral patterns of 3D objects in a known scene. Therefore, the most relevant information required from a 3D object recognition and tracking system is the 3D dynamic positions, orientations and occupancies, i.e. volumes, of all the moving objects and objects that can move. Detailed information about the surface shape of individual objects is not of great concern or relevance here. This has important implications in determining the type of object representation and corresponding image descriptions.

Object recognition is one of the key visual tasks in the interpretation of dynamic activities captured in image sequences. One of the most widely used volumetric model representations is the generalised cylinder [7]. The advantages of volumetric object models are that their global properties (volumes, positions and orientations) are directly represented and easy to obtain and only a small set of values are needed to parameterise them. In contrast, surface-based representations are based on piece-wise reconstruction of object surfaces, planar or curved. The surface reconstruction approach concentrates on recovering detailed information about the local geometric shape of an object. This is important in recognising objects that are primarily distinguished by their differences in shape and essential for object manipulations such as those required in robotics. However, this approach is computationally expensive and it is difficult to access global properties of objects since they are not represented and need to be inferred from the local shape information. A volumetric representation scheme, then, seems to be most appropriate for behavioral evaluation in surveillance because conceptual descriptions will need to be recovered on-line from the global properties of the objects over time.

In principle, the choice of model representation will determine the type of symbolic image description, i.e. specific extracted image features, so that they can be used to match effectively with projected geometric features on object models. The essential nature of the matching process is that the mapping between positions of the image features and the position and orientation of the models is given by a set of non-linear functions. Although there are many model-matching techniques, it appears to us that the techniques developed so far have not adequately resolved the issue of consistent object matching in cluttered and fast changing scenes, especially the problem of invoking the right model. We propose to avoid this problem by starting the evolution of a dynamic interpretation using the kind of simple generic volume model described above with the parameters for spatial extent, position on the ground-plane and motion refined over time by the visual evidence.

2.4 Tracking Dynamic Objects

Model-based object recognition techniques have been adopted as one of the key components in surveillance systems [24, 45]. However, it is recognised that the temporal correlations between objects over time have not been fully incorporated into the recognition process. In other words, although model-based object tracking is generally required in the understanding of a dynamic scene with moving objects, most of the proposed schemes only address the problems of matching static 2D image descriptions to 3D object models over time. For example, Worrall *et al.* [45] used direct matching of image descriptions to projected descriptions of object shape, position and orientation in every frame. Others [24, 28] have advanced the approach by applying an independent closed form motion model for each object and match the detected static image descriptions with the motion model in each frame in order to optimise the predicted motion parameters.

However, the essence of surveillance is being able to detect and interpret change in the scene. Tracking schemes that ignore the available information about temporal changes in the image are likely to complicate the interpretation tasks in the later stages of the processing. Measuring image motion not only leads

to a more compact representation of an image sequence over time, it also distinguishes noise and possible objects of interest in the scene. In general, it is well understood that the dense image motion (optic flow field) contains valuable information not only about 3D shapes of the scene but also motions in the scene. For surveillance in particular, image motion alone can be sufficient for providing information for statistical interpretations such as measuring traffic density on roads or passenger population on train platforms. It also provides one of the most obvious bootstrapping cues for initialising object model matching when needed.

The measurement of image motion is the primary cue for detecting movement in the early stages of the interpretation process and has to be computed for effective interpretation of dynamic scenes. There have been many approaches to interpreting image motion, for example [9, 26], but there has been little attempt to map dynamic image descriptions such as the optic flow field to any global 3D descriptions of objects in a model-based object recognition scheme. Only the highly computationally expensive option of detailed reconstruction of 3D surface and edges and subsequent detailed geometrical matching has been explored [29].

Most model-based object tracking techniques have adopted a simplified motion model based on the Extended Kalman Filter (EKF) in updating an object's 3D motion parameters. However, the difficulties in getting a good initial estimation of motion, which is essential to allow meaningful predictions, have again been mostly overlooked. A new probabilistic relaxation framework that reflects some of the Bayesian belief revision principles has been proposed by Kittler [23] and illustrated for matching relational structures. This works by mapping evidence to expectations between different representation domains such that the most likely interpretations are obtained. This reinforces our idea that in order to overcome the limitation of the closed form EKF approach in motion estimation, a distributed belief revision approach that incorporates probability evaluations with non-linear constraint satisfaction networks is required. It would be more appropriate for this matching problem since much weaker constraints between the evidence and expectation are imposed.

2.5 Behavioural Representation

To compute behavioural descriptions, a simple notion of time and events is required in an on-line vision system. Some researchers, for example Brooks [8], have even suggested that we can dispense with all internal representations and allow the world to be its only model using reactive control for intelligent activity. However, we propose to go only part way towards this approach with a limited situated analysis in which we represent the properties that are relevant for our visual tasks. These properties will enable us to identify when a change occurs in a meaningful context. For example in surveillance, this context is the representation of our ground-plane (scene) knowledge and other current purposively moving objects. Nagel [30] reviewed the few projects which have tackled the problem of delivering conceptual descriptions in road traffic domains. These include NAOS [31] and CITYTOUR [33] which allow question-answering as an off-line query process. Nagel also considers the problem of on-line generation of such descriptions. His approach goes some way towards the goal of specifying conceptual descriptions in terms of motion verbs that could be effectively computed. However, we think a more situated approach has advantages as it uses a perceiver-centred or "deictic" frame of reference. For example, spatial deixis uses words like "here" and "there" and temporal deixis uses "yesterday" and "now". Deictic reference depends on knowledge of the context but can decompose and simplify the reasoning that needs to be done compared to the global "state-based" approach of traditional AI. It seems deeply embedded in our communication and spatial reasoning [33].

We require a more local frame in surveillance than that required to support full cognitive planning or coordinating actions as we are only observing the activity of moving objects. Formal logic approaches using well-defined languages with clear meaning for time, events, and causality, for example [2, 35], are useful for validating and prototyping new approaches to behavioural analysis but do not seem suitable for on-line vision systems. To effectively mix both qualitative and quantitative descriptions of space and time for a wide set of visual tasks, we used Fleck's cellular models [15, 16] which discretise real time and space. In cellwise-time, as developed by Fleck, each cell is a state and we can classify change as either: "state-changes", which involve sharp change; "activities", where continuous change occurs; or "accomplishments", which are composites of activity and state-change. "Episodes" are seen as composed of a starting state-change, an activity, and an end state-change in this framework. This kind of representation can be called "analogical" and has been further developed for the representation of events and behaviour under task-based control for surveillance [21, 22].

3 Bayesian Techniques

To implement our active behavioural analysis, we use the Bayesian belief revision approach which is conceptually attractive and computationally feasible for vision [25, 34]. In an on-line system simple correlations in the spatio-temporal data can be exploited to efficiently infer quite complex behaviour. Modelling and updating the dependent relationships and their probability distributions in belief nets is relatively easy both off-line or on-line. We have demonstrated this for both motion segmentation and tracking [19] and in the evaluation of visual behaviour [22]. Conceptually, we are addressing the issue of modelling an information retrieval process that purposively collects evidence in the image to support interpretations of dynamic behaviours in the scene. Ambiguity in the interpretation of individual levels of computation means that context-dependent information integration is required to obtain more coherent interpretations of the visual evidence. Recent developments in probabilistic relaxation, belief and decision theory have provided us with

a sound computational base [32] which we can extend for the problem at hand.

Bayesian belief networks are Directed Acyclic Graphs (DAGs) in which each node represents an uncertain quantity using variables. The arcs connecting the nodes signify the direct causal influences between the linked variables with the strengths of these influences quantified by associated conditional probabilities. We use only singly connected trees for modelling as these are fast to update. A selection of these multivalued variables will be the direct causes (parents) at a particular node. The strengths of these direct influences are quantified by assigning a link matrix for every combination of values of the parent set. The conjunction of all the local link matrices of variables in the network specifies a complete and consistent global model which is given by the overall joint distribution function over the variable values. The behaviour of a visual process is partially defined by its processing parameters which are updated in the network so dynamic evaluation will be consistent with the visual task.

In a belief network, we can quantify the degree of coherence between the expectations and the evidence by a measure of local belief and define belief commitments as the tentative acceptance of a subset of hypotheses that together constitute a most satisfactory explanation of the evidence at hand. Bayesian belief revision updates belief commitments by distributed local message passing operations. Instead of associating a belief measure with each individual hypothesis, belief revision identifies a composite set of hypotheses that best explains the evidence. We call such a set the Most-Probable-Explanation (MPE). The conceptual basis of the propagation mechanism that updates the network is quite simple. For each hypothetical value of a single variable, there exists a best extension of the complementary variables. The problem of finding the best extension can be decomposed into finding the best complementary extension to each of the neighbouring variables according to their conditional dependencies. This information can then be used to decide the best value at the node. The decomposition allows the processing to be applied recursively until it reaches the network boundary where evidence variables have predetermined values.

4 Experimental Results

4.1 Motion Segmentation and Tracking

In VIEWS, one of the key objectives is to segment detected optic flow field into dynamic regions corresponding to possible moving objects and to track these regions effectively and consistently over time. Wenz [43] applied a scheme based on estimated frame displacements of the extremal loci of a bandpass filter. Similar displacement vectors are grouped into different moving regions (bounding boxes) in each frame and the similarity is defined by four parameters 1) neighbourhood range, 2) neighbourhood displacement magnitude ratio, 3) neighbourhood orientation difference and 4) neighbourhood vector numbers. In this direct approach, these similarity parameters are set as independent constants across the entire image for computational simplicity. However, it is unable to deliver consistent interpretations in images of crowded scenes such as the traffic roundabout shown in the top picture of figure 3. The example frames in figures 6, 7 and 8 illustrate some typical defects in the sensitivity and consistency of the direct approach. We used scene-oriented contextual knowledge in the control of parameter values to overcome these problems.

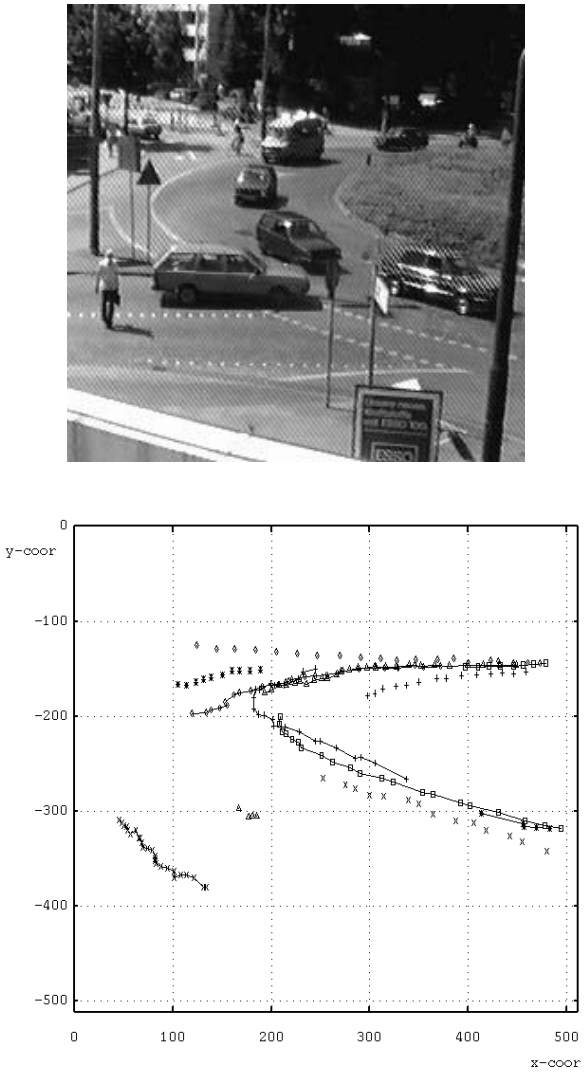

Figure 3: *Top: a traffic roundabout scenario and its traffic flow. Bottom: correlated spatio-temporal constraints on the movements of individual objects are imposed implicitly by this scene layout.*

VIEWS uses a fixed camera for collecting visual input in each scenario. Under such static camera configurations, the three dimensional scene layout imposes indirect, but nevertheless invariant, constraints on both possible loci of appearances, sizes, speeds of bounding boxes and the overall traffic flow. The bottom picture of figure 3 illustrates the recorded mo-

tion patterns of vehicles on the roundabout over 450 frames. The *a priori* constraints on object size, speed and relationships between the parameters in the interpretation can be analysed and used to initialise the probabilities in a Bayesian network. The following correlated measures (with respect to image coordinates) are constrained probabilistically in the parameter net: 1) between object orientation and optic flow vector orientation; 2) between object size and flow vector neighbouring speed ratio; 3) between neighbouring orientation difference, object dx, object dy and object bounding box width or height. Such probabilistic constraints on the bounding boxes set up a compound network of coherent hypotheses (figure 4) that is modelled by a Bayesian belief network with dynamic belief revision propagation.

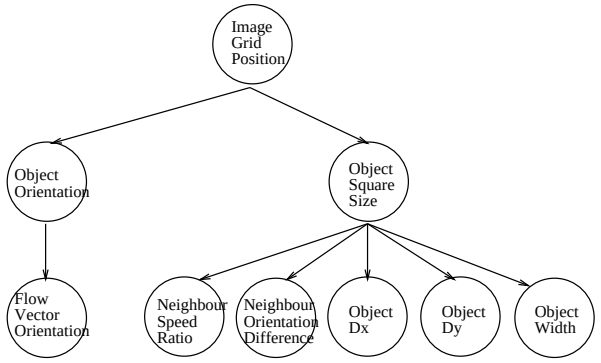

Figure 4: *A belief network that captures the dependent relationships between the scene layout and relevant measures in motion segmentation and tracking.*

The belief network in figure 4 has a tree structure, a special type of "singly connected" network, in order to guarantee the propagation of message passing in belief revision is tractable [32]. The image is divided into grids and the root node of the tree IGP (Image Grid Position) represents the probabilistic expectation of occurrence for objects in each image grid position. Nodes OSS and OOR represent respectively the probabilistic expectations in the square size and orientation of bounding boxes in image grids. The six leaf nodes at the bottom level of the tree represent, respectively, the expectations in flow vector orientation (FVO), neighbouring vector speed ratio (NSR), orientation difference (NOD), x component in object bounding box's displacement (ODX), y component in bounding box displacement (ODY), and the width of a bounding box (OWD).

It is important to point out that first, leaf nodes are the evidence nodes and it is desirable to relate them to qualitative measures by representing relative measures between flow vectors. This is designed to overcome the instability of individual vectors in optic flow fields. Second, great effort was made to reduce the number of causal connections and the number of hypothetical variables to the minimum at the expense of approximations in the representation of certain variable nodes. This is because the computational load increases by an order of $2^n - 1$ where n is the number of variable nodes in a network [11]. Third, in order to have efficient computation, it is useful to approximate any continuous variable with a set of few discrete values. Fourth, the conditional probability distribution matrices between any two nodes are usually subject to probabilistic estimation based on extensive test examples. Statistical studies in the past [11] suggest that if a well controlled number of variables are built into a Bayesian network, the estimated distribution matrices capture the general characteristics of the problem. Accurate estimation of these parameters remains one of the important factors for the computational success of a belief network. Recent studies by Spiegelhalter [36] have shown techniques for updating and learning the distribution matrices dynamically in order to provide more accuracy in their estimation.

The current design of the belief network has been tested extensively on image sequences from the traffic roundabout scenario. The sensitivity of the techniques is measured by their "false alarm rate", which was taken over an image sequence of 400 frames using a strict criterion of matched "true" (identified by human visual analysis on a frame by frame basis) and the automatically computed bounding boxes. The top graph in figure 5 shows the false alarm rate on both techniques over time. It gives a good indication that the belief revision approach increases true identifications significantly without introducing excessive false alarms. Throughout the whole sequence, the maximum false alarm rate from the belief revision approach is about 16 %, which is below the minimum rate from the direct approach. The maximum false alarm rate of the direct approach for this case, on the other hand, reaches 60 % and its average rate is nearly 50 %.

To obtain the consistency measures, we compiled the histories of tracked objects from both techniques and compared them with the "ground truth" from a 170 frame image sequence. In the bottom graph of figure 5, the flattest line shows the ground truth of the number of objects against their durations in the scene. For example, 1 object stayed for the entire 170 frames, 13 objects lasted for 14 frames, etc. The direct approach fragmented objects with long durations and tracked them as a large number of objects with very short histories. No object is tracked for more than 50 frames, which is the basis of the poor consistency of the direct approach. In contrast, the belief revision approach provides us with more accurate measures of both the number of objects and their durations.

To estimate the computational cost, we measure the time consumption (in seconds) of both schemes over the 400 frame sequence. The frame by frame computational overhead throughout the whole sequence is below 13 %, and it is worth pointing out that providing more accurate segmentation and tracking of objects instead of missing identifications will always require "extra" computation.

The quantitative measures presented here illustrate that: with very limited cost in computational time, significant gains are obtained in effectiveness and consistency by using the belief revision technique. A more visual comparison between the two approaches can be

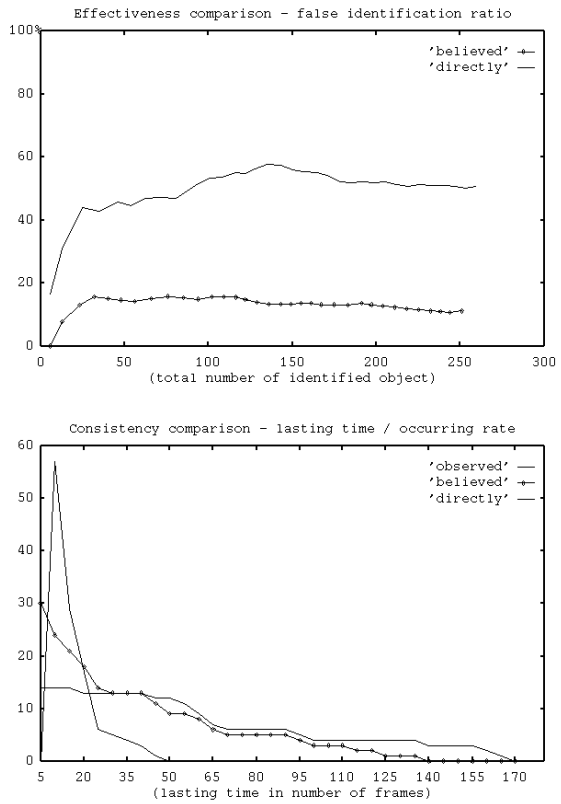

Figure 5: *Top: the false alarm rate. Bottom: the "ground truth" and detected number of objects and their duration in the scene.*

seen in figures 6, 7 and 8.

Three successive frames from our test sequence are shown with the results from the direct approach at the top and from the belief revision approach on the bottom. It is worth noticing that: first, the belief revision approach is very robust against incomplete evidence (see the tracked cyclist behind a sign post to the left hand side of frames 145 and 150). Second, it is capable of segmenting very close moving objects (see the cyclist and the two cars close to its right). Third, in these examples, it consistently identifies all the moving objects. Note that these are quite cluttered traffic scenes and it is typical of these techniques to show more sensitive and robust performance on "difficult" scenes. For simple cases, direct methods can work well but have clear limitations for the full set of cases met in real world video sequences.

4.2 Behavioural Evaluation

Bayesian techniques can also be used to overcome the problem of uncertainty and incompleteness in the evaluation of behaviour by bringing both task-based and scene-based knowledge into the interpretation process. This behavioural interpretation requires both modelling what one is looking for (top-down expectations) and interpreting evidence of what could be appearing (bottom-up inference). The prior probabilities can be used to initialise the network and then the evidence is dynamically interpreted under the current expectations using both top-down (λ messages) and bottom-up(π messages) updating of values in the network. We have effectively used such Bayesian networks together with a deictic representation both to create a dynamic structure to reflect the spatial organisation of the data and to measure task relatedness [22]. We integrate the behavioral evaluation and interpretation by giving a combined attentional focus for the road traffic examplar where the behaviours of interest were "overtaking", "following", "queueing" and "unknown". For example, a simple proximity cue invokes the behavioral analysis of overtaking. A task-based Bayesian network (adapted from [34]) is used in modelling spatial and temporal relationships in order to direct the evidence collection in the image sequence.

Figure 6: *Frame* 140*. Top: direct approach fragments object bounding boxes (compare with next two frames). Bottom: belief revision approach captures most moving objects consistently in successive frames.*

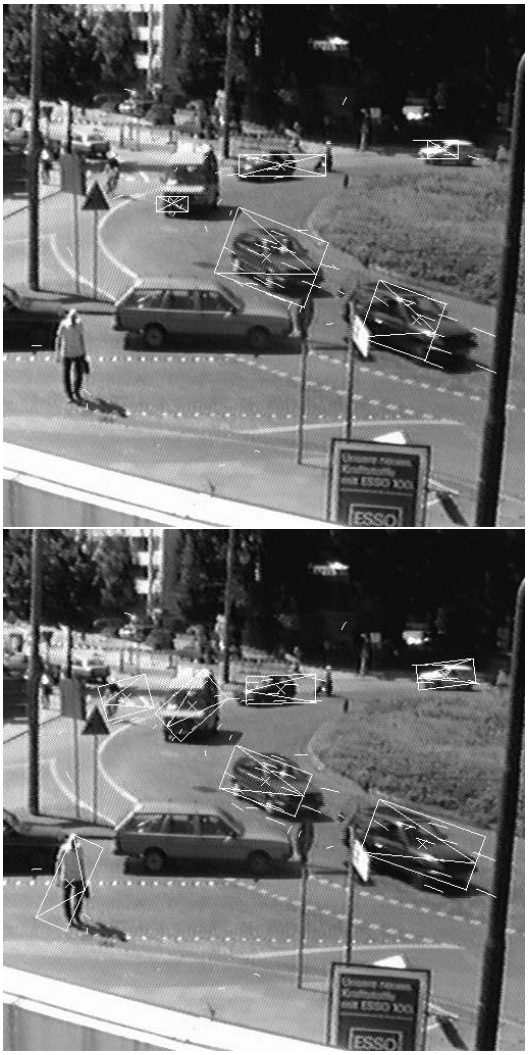

Figure 7: *Results on frame* 145. *Top: by the direct approach. Bottom: by the belief revision approach.*

Figure 8: *Results on frame* 150. *Top: by the direct approach. Bottom: by the belief revision approach.*

We use a separation of preattentive (peripheral) and attentional processing in our behavioural evaluation system. The simple peripheral operators such as velocity and proximity on the ground-plane act as cues for the more complex attentional operations such as path-prediction and computing deictic spatial relationships which are used in the full evaluation. An attentional mechanism guides the application of appropriate complex evaluation in a particular dynamic context and makes the reasoning relevant to the current task. This attentional mechanism uses an agent-based formalism implemented by Bayesian network updating. The objects in our scenes have an "intrinsic front" which defines each object's frame of reference in the deictic representation. We have developed "typical-object-models" for the interpretation of behaviour using time-ordered combinations of deictic relationships between objects of interest in particular ground-plane contexts. The deictic relationships may be simply ("behind" "previous"), ("beside" "now"), and ("infront" "next"). The simple peripheral operators are applied to all our segmented, tracked objects and the typical-object-model determines the specific attentional operations to be performed. It also determines which values should be saved and the set of operations to be performed on the next clock tick. The results are fed back to the appropriate agent to give task-related features for future selection. The approach here is related to [1] but extended to deal with several local deictic viewpoints. In this way an agent need not describe every object in the domain but only those relevant to its particular task.

To combine the information that develops over time we use a dynamic form of Bayesian network (DBN) which captures the changing relationships between scene objects. The DBN is composed of temporally separated subgraphs that are interconnected using reconfigurable links. The node-building and node-

linking updates the structure for the current time so that a new tree is obtained which inherits values for beliefs at nodes referenced in the "official-observer" view. These markers, maintained centrally by the official-observer, solve the problem of consistently associating entities which are maintained only locally by the distributed agents. A matrix of conditional probabilities captures the interest of the proximity relationships according to whether the objects are: "not-near", "nearby", "close", "very-close" or "touching". If, at the current time point, A and B remain near each other, we form the temporal links and extract a new tree structure rather than a multiply connected network. In the example here, each tree only holds values from the current and immediately previous time points and a tree is formed for each relationship so it is trivial to prove that no loops are introduced, which is important for bounding the computation. Once all propagations are complete, the most interesting relationship can be selected.

The structural changes in the DBN reflect the simple, monitored relationships between all objects of interest. For those objects involved in the most interesting relationships, this DBN is augmented by a "tasknet", which is a structurally fixed Bayesian network that builds a coherent interpretation of the temporally evolving behaviour. The input nodes here represent key features relevant to the task it has been constructed to identify and the output root node represents the overall belief in the behavioural task based on evidence collected so far. Allocating an attentional process explicitly in this way ensures that a tasknet is running on the selected objects and can terminate when an uninteresting situation is recognised.

The attentional system has three properties: "focus-of-attention" which ensures only the target hypothesis and associated functions are updated; "terminate-attention" which can stop all activities associated with a target hypothesis that has been confirmed or denied to an acceptable degree of confidence; and "selective-attention" which allows dynamic selection of the most interesting hypothesis to watch. We can only attend a limited number of objects so a measure of "utility" is computed from the cost of performing the processing and its "interestingness". This determines which objects will be "watched". The overall effect of the attention and control here is the formation of behavioural descriptions that evolve over time to capture what is happening in the scene rather than being dependent on observing the whole episode before anything can be said to have happened.

5 Conclusion and Future Work

To summarise the arguments in the earlier sections, we are suggesting that advanced visual surveillance can benefit from taking a purposive approach to system design. This means using representations that are closely tailored to the surveillance tasks. We also suggest using a purposive approach in the system framework so that task-dependent processing under an active focus of attention can selectively gather evidence under the current set of expectations. In the design of a surveillance system, we point out that off-line processing, such as camera calibration or setting up of the ground-plane representation, can afford elaborate models to provide the required accuracy and functionality. However, we argue that there is the need to find the simplest possible models of object structure and behaviour in the on-line processing to ensure both fast interpretation times and robustness with generality. For visual surveillance, we argue that the basis for the initial detection of moving objects and cues for object position and velocity should be simple visual motion measures. We also argue that the high-level interpretation, in general, requires behavioural models that are decomposable into simple primitives that can be detected in real-time and that the evolving behavioural descriptions should be computed under context-based expectations. The intermediate processing, however, poses more problems as existing model-based tracking techniques are designed for a small set of detailed models with limited dynamic updating. For more demanding surveillance tasks where dynamic scene and event discrimination is the key, we propose the formulation of a new scheme within the Bayesian belief network framework as we have argued that this will provide the kind of "weak" combination of constraints appropriate for incremental shape and motion recovery in the face of uncertain and incomplete visual evidence. The experiments described above illustrate the feasibility and computational tractability of this approach.

In section 2, we analysed the components required in our advanced surveillance systems. We briefly reviewed progress so far and recommended the particular models and associated processing schemes that show the most promise. For example, we suggested a full radial alignment model for a fixed wide-angle surveillance camera in off-line calibration. However, if we require on-line dynamic surveillance, we would suggest simplifying the model and extending the Bayesian networks to provide a coherent model right down to the level of camera control. In the spatial and behavioural representations, we proposed cellular models as they support both qualitative and quantitative aspects of the processing as required. These models provide fast contextual indexing of computational constraints in the behavioural analysis. They have been integrated into the Bayesian belief networks to provide a framework in which interpretation evolves dynamically with a task-dependent focus of attention. In the image and object representations, we suggested using simple, reliable measures of visual motion together with volumetric models that give immediate access to the global properties of position, orientation and ground-plane motion which are required for the behavioural analysis of the moving objects. These competances, then, were all derived from a purposive strategy in visual system design.

In section 3, we outlined Bayesian network techniques. These provide a means of performing both bottom-up, data driven processing and top-down, expectation driven processing in the on-line computations. Bayesian nets allow the computation of the Most-Probable-Explanation of visual evidence under the expectations at all levels of abstraction in a vi-

sion system. The nodes in the parameter network are abstract entities that can be associated, for example, with simple low-level interpretations of the position and speed of bounding-boxes corresponding to possible moving objects in the image-plane. They can equally well be associated with high-level interpretations of the kind of behaviour in which a particular object is engaged. The network updating techniques implement a fast non-recurrent solution using the current values of the nodes. The updating rules were derived from Bayesian theory which makes them very well suited to the analysis of essential visual changes in surveillance where there is always a great deal of uncertainty and incompleteness in the data. It is also important to note that the Bayesian networks allow for task-based control which is required to make the processing performed by the system selective and avoid the combinatorial explosion entailed in passive analysis. It is still important, however, to keep the networks as simple as possible and model only the essential dynamic dependencies; those that allow a rapid evaluation of the evolving spatio-temporal patterns of behaviour.

In section 4, another aspect of the Bayesian networks becomes apparent, the ability to encode knowledge by modelling dynamic dependencies amongst the visual parameters through examples and prior probabilities of classes of interpretation. This is possible by analysis of the problem where there are obvious scene-based constraints such as traffic flow direction in certain lanes of a roundabout. It is also possible to learn these constraints and dependencies using appropriate techniques [36, 44]. This can be a time consuming process but is typically computed off-line with only limited adaptive refinement on-line. The requirement to turn conceptual scene-based or task-based knowledge into a readily accessible form for real-time processing has been recognised in the past. Many hybrid schemes using both knowledge-based and numerical techniques have been proposed but would not easily support real-time systems. On the other hand, however, the kind of constraints that can be imposed using numerical techniques are rather inflexible for advanced visual surveillance where we have a lot of domain specific knowledge. We would also argue that neural network approaches would be very difficult to develop for this class of applications. Nor do we think it possible to "evolve" solutions to such complex problems using genetic algorithms. It thus seems to us that the most promising unified framework is provided by Bayesian belief revision networks as we have sucessfully demonstrated for a set of typical advanced surveillance tasks.

Acknowledgements

We are grateful to Richard Howarth for some of the figures used in this paper and would also like to thank him for valuable comments on the initial draft.

References

[1] P.E. Agre and D. Chapman. "Pengi: An implementation of a theory of activity". In *AAAI National Conference on Artificial Intelligence*, 1987.

[2] J.F. Allen. "Towards a general theory of action and time". *Artificial Intelligence*, 23, 1984.

[3] Y. Aloimonos, I. Weiss, and A. Bandopadhay. "Active vision". In *International Conference on Computer Vision*, London, England, 1987.

[4] R. Bajcsy and P. Allen. "Sensing Strategies". In *US - France Robotics Workshop*, 1984.

[5] D. Ballard. "Animate vision". *Artificial Intelligence*, 48, 1991.

[6] T.O. Binford, T.S. Levitt, and W.B. Mann. "Bayesian inference in model-based machine vision". In *Uncertainty in Artificial Intelligence 3*, North-Holland, 1989.

[7] R.A. Brooks. "Symbolic reasoning among 3D models and 2D images". *Artificial Intelligence*, 17, 1981.

[8] R.A. Brooks. "Intelligence without reason". In *International Joint Conference on Artificial Intelligence*, Sydney, Australia, 1991.

[9] B.F. Buxton, D.W. Murray, H. Buxton, and N.S. Williams. "Structure-from-motion algorithm for computer vision on an SIMD architecture". *Computer Physics Communications*, 37, 1985.

[10] H. Buxton and N. Walker. "Query-based visual analysis: Spatio-temporal reasoning in computer vision". *Image and Vision Computing*, 6, 1988.

[11] E. Charniak. "Bayesian networks without tears". *AI Magazine*, 12(4), 1991.

[12] VIEWS consortium. "The VIEWS project and wide-area surveillance". In *ESPRIT Workshop at ECCV*, Genoa, Italy, 1992.

[13] D.R. Corrall, A.N. Clark, and A.H. Hill. "Airside ground movements surveillance". In *NATO AGARD Symposium on Machine Intelligence in Air Traffic Management*, Berlin, Germany, 1993.

[14] E.D. Dickmanns. "A general dynamic vision architecture for UGV and UAV". *Journal of Applied Intelligence*, 2, 1992.

[15] M. Fleck. "Representing space for practical reasoning". *Image and Vision Computing*, 6, 1986.

[16] M. Fleck. *Boundaries and Topological Algorithms*. PhD thesis, MIT AI Lab., 1988.

[17] S.G. Gong. "Visual observation as reactive learning". In *International Conference on Adaptive and Learning Systems*, Orlando, Florida., 1992.

[18] S.G. Gong and H. Buxton. "On the expectations of moving objects". In *European Conference on Artificial Intelligence*, Vienna, Austria, 1992.

[19] S.G. Gong and H. Buxton. "Bayesian nets for mapping contextual knowledge to computational constraints". In *British Machine Vision Conference*, Guildford, England, 1993.

[20] R. Howarth. *Spatial Representation, Reasoning and Control for Visual Surveillance.* PhD thesis, QMW, University of London, 1994.

[21] R. Howarth and H. Buxton. "An analogical representation of space and time". *Image and Vision Computing*, 10, 1992.

[22] R. Howarth and H. Buxton. "Selective Attention in Dynamic Vision". In *International Joint Conference on Artificial Intelligence*, Chambery, France, 1993.

[23] J. Kittler, W.J. Christmas, and M. Petrou. "Probabilistic relaxation for matching problems in computer vision". In *International Conference on Computer Vision*, Berlin, Germany, 1993.

[24] D. Koller, K. Daniilidis, T. Thorhallson, and H.-H. Nagel. "Model-based Object Tracking in Traffic Scenes". In *European Conference on Computer Vision*, Genoa, Italy, 1992.

[25] T.S. Levitt, T.O. Binford, and G.J. Ettinger. "Utility-based control for computer vision". In *Uncertainty in Artificial Intelligence 4*, North-Holland, 1990.

[26] H.C. Longuet-Higgins and K.F. Prazdny. "The interpretation of a moving retinal image". *Proc. Royal Society of London*, B-208, 1980.

[27] D. Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information.* W.H.Freeman & Co., 1982.

[28] R. Marslin, G.D. Sullivan, and K.D. Baker. "Kalman filters in constrained model-based tracking". In *Proceedings of the British Machine Vision Conference*, Glasgow, Scotland, 1991.

[29] D.W. Murray, D.A. Castelow, and B.F. Buxton. "From image sequences to recognised moving polyhedral objects". *International Journal of Computer Vision*, 3, 1988.

[30] H.H. Nagel. "From image sequences towards conceptual descriptions". In *Image and Vision Computing*, 6, 1988.

[31] B. Neumann. "Natural language description of time varying scenes". In *Semantic Structures.* Lawrence Erlbaum Associates, 1989.

[32] J. Pearl. *Probabilistic Reasoning in Intelligent Systems, Networks of Plausible Inference.* Morgan Kaufmann, 1988.

[33] G. Retz-Schmidt. "Various views on spatial prepositions". *AI Magazine*, 9(2), 1988.

[34] R.D. Rimey and C.M. Brown. "Where to look next using a Bayes net: Incorperating geometric relations". In *European Conference on Computer Vision*, Genoa, Italy, 1992.

[35] Y. Shoham. *Reasoning about Change: Time and Causation from the Standpoint of Artificial Intelligence.* MIT Press, 1988.

[36] D.J. Spiegelhalter and R.G. Cowell. "Learning in probabilistic expert systems". In *Bayesian Statistics 4.* Oxford University Press, 1992.

[37] A. Toal and H. Buxton. "Spatio-temporal reasoning within a traffic surveillance system". In *European Conference on Computer Vision*, Genoa, Italy, 1992.

[38] R.Y. Tsai. "A versatile camera calibration technique using off-the-shelf TV cameras and lenses". *Journal of Robotics and Automation*, RA-, 1987.

[39] J.K. Tsotsos. "Knowledge Organisation and its Role in Representation and Interpretation for Time-Varying Data: the ALVEN System". *Computing Intelligence*, 1, 1985.

[40] J.K. Tsotsos. "On the relative complexity of active vs. passive visual search". *International Journal of Computer Vision*, 7, 1992.

[41] S. Ullman. "Visual routines". *Cognition*, 18, 1984.

[42] J. Weng, P. Cohen, and M. Herniou. "Camera calibration with distortion models and accuracy evaluation". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14, 1992.

[43] G.H. Wenz. *Parallel Realtime Detection and Tracking.* MSc thesis, QMW, University of London, 1994.

[44] S.D. Whitehead and D.H. Ballard. "Active perception and reinforcement learning". *Machine Learning*, 7, 1991.

[45] A.D. Worrall, R.F. Marslin, G.D. Sullivan, and K.D. Baker. "Model-based tracking". In *Proceedings of the British Machine Vision Conference*, Glasgow, Scotland, 1991.