Visual Interpretation and Understanding
CSRP 452

Hilary Buxton
School of Cognitive and Computing Sciences
Falmer, Brighton BN1 9QH
hilaryb@cogs.susx.ac.uk

January 1997

Abstract

In this review, the subfield of visual interpretation and understanding is first defined and three major issues for using knowledge to increase the functionality and performance of vision systems are introduced. These selected issues concern the role of context, control and learning. In section 2, four approaches to reasoning are distinguished and illustrated with key papers on 1) constraint-based vision, 2) model-based vision, 3) formal logic, and 4) probabilistic frameworks for visual interpretation and control. In section 3, exploitation of these techniques is discussed for automating linguistic descriptions of scenes, enhancing human computer interaction in multimodal and multimedia systems, in behavioural control for robotics, advanced surveillance systems and biomedical image analysis systems. Finally, promising directions for future research are suggested. These use the deformable models, dynamic learning, and situated approaches to visual understanding discussed in the main sections of the report.

1 Introduction

Most current research on computer vision attempts to model generic capabilities for image feature detection and region segmentation, stereo and motion perception, object recognition and tracking etc. However, these general approaches are not sufficient, in themselves, to cope with the wide variability in real-world scenes and task-specific requirements of many applied vision systems. The subfield of visual interpretation and understanding combines techniques from AI and knowledge-based systems with computer vision techniques to deliver enhanced functionality in such systems. Naturally, we encounter many of the major issues in AI such as knowledge representation and reasoning, control and the handling of uncertainty, as well as machine learning. Much of this work assumes that knowledge drives reasoning in visual interpretation, using expectation or a hypothesis to direct the processing. This means we take sides in one of the great debates in vision research since knowledge-based vision introduces a major "seeing as" bias, rather than being just another level of generic processing. In our subfield, visual context is seen as essential for understanding what is depicted in images or image sequences. If we are to build efficient systems that can tackle many different tasks, high-level attention and control of the processing is also seen as essential. In addition, if we are to incorporate scene and task knowledge, we have to address the question of how such knowledge can be acquired. Formal structural and propositional knowledge has to be designed by hand but, as we will see, some representations can be learnt.

Research in this subfield goes beyond the recognition of features and objects to give descriptions of the scene content that are meaningful to the observer or user of the system. To achieve this, we build in domain specific knowledge of the scene and tasks by *representing prior knowledge* in a readily accessible form. For example, in VIEWS [12, 18], which was a major European knowledge-based vision project, advanced visual surveillance capabilities were developed using a mixture of constraint-based, model-based, logic-based and probabilistic interpretation techniques. Demonstrations showed that, for object detection and tracking, performance was much improved by scene-based knowledge of expected object trajectories, size and speed [26, 27]. Also, both scene and task-based knowledge allowed selective processing under attentional control for behavioural evaluation in traffic scenes [31, 32]. There are many such applications where the existence of prior scene and task knowledge provides context for conditioning computer vision algorithms.

1.1 Context

Traditional computer vision systems attempt to recognise static objects and their dynamic behaviour using bottom-up, data-driven processing with minimal use of prior knowledge. However, such systems are bound to fail for complex domains as the information in the images alone is insufficient for detailed interpretation or understanding of the objects and events. Knowledge-based vision research relies primarily on scene context to overcome this kind of uncertainty. For example, Strat and Fischler [60, 61] combine many simple vision procedures that analyse colour, stereo, and range images with relevant contextual knowledge to achieve reliable recognition. There are many other types of contextual knowledge such as functional context [59], where attributes such as shape are used to infer the functional role of the object and direct the visual processing [6]. Another type of context, which is particularly relevant to multimodal and multimedia systems, is linguistic context [55, 62]. In addition, task context is an important source of control for the visual processing [13, 20]. The role of context, then, is central to visual interpretation and understanding and representing context in an appropriate way, so that it improves the effectiveness and efficiency of visual reasoning, is a key issue in the field.

1.2 Control

Another failing of traditional vision systems is the lack of attention given to purposeful, selective control of the processing. This is again a key issue when real-time, dynamic applications are being developed. Ballard's influential paper [2] signalled the start of a new consensus in the computer vision community that vision is active, highly selective and purposeful. Rao and Ballard [49] have also proposed an active vision architecture, inspired by the organisation of human visual processing, that uses simply acquired and indexable iconic representations. Interdisciplinary research also shows that the task and the nature of the scene determine visual attention and can allow selective tuning of visual processing [68]. This requirement for highly selective visual processing was the main theme of VAP [19, 20], which was a major European project to develop active visual processing. Much of this active vision research has concentrated on camera control, navigation and lower-level visual tasks which do not involve visual understanding using stored knowledge. However, ideas from active vision have been extended and applied to high-level reasoning [13, 33]. This control of visual processing is deeply task-dependent and usually requires indexable knowledge structures for real-time systems.

1.3 Learning

A further development in this subfield involves representations to support task-level control and learning. For example, by using Hidden Markov

Models (HMMs) we can used for learning probabilistic relationships for eye movement control [51] and applied to modelling of vehicle trajectories [26]. On-line updating using such visually augmented HMMs enables both tracking and reporting of these purposive vehicle movements. More recently, Bayesian Belief Networks (BBNs) have been used to support the learning of both initial and conditional probabilities for camera control [52] and for segmenting and tracking vehicles [27]. In addition, BBNs have been used with behavioural models to provide task-dependent control in behavioural analysis [15, 33]. These kinds of learning are essentially conditional parameter estimation using the statistics of example image sequences. Learning dynamic, parametric models for visual motion patterns [7] is an important capability for intelligent tracking. Also, learning statistically-based deformable models is crucial for many medical applications [17], and in tracking moving people [3] for surveillance. In these examples, the knowledge is acquired off-line and exploited in the on-line system. The role of *learning* can be extended to behavioural models by using on-line evaluation or reinforcement learning [53, 69] in order to create a more open system that can adapt its behaviour to the changing environment. This is an exciting field in which we can envisage fully autonomous visual agents learning their own goals and representations.

2 Approaches

Reasoning is the main focus of work in visual interpretation and understanding so here we discuss four major approaches 1) constraint-based vision, 2) model-based vision, 3) formal logic, and 4) probabilistic frameworks. In each of these subsections, we first describe the general history of the approach and then go on to recent developments.

2.1 Constraint-based Reasoning

VISIONS (Visual Integration by Semantic Interpretation of Natural Scenes) [28] was an early knowledge-based system for static image interpretation. An early example for dynamic scene analysis, which also used constraint-based reasoning, was the ALVEN system [67]. In contraint-based vision, a set of interacting constraints about the scene and task context are used to guide the reasoning. For example, the VISIONS system had many levels for representation in both the long-term knowledge base and the short-term interpretation of a particular image. It used both declarative and procedural knowledge in hypothesis generation using bottom-up and top-down reasoning. The schema mechanism supported a conceptual hierarchy by allowing entities to be described as themselves, part of a higher level schema, or a schema for lower level entities. It was necessary to develop VISIONS to incorporate Bayesian belief probabilities [48] and more recently, Dempster-Shafer belief functions [24] to handle uncertainty in visual evidence. This move from simple constraint-based reasoning to incorporate more sophisticated probabilistic reasoning with the symbolic knowledge has been one of the major trends in research on visual interpretation and understanding. This is mainly because it allows more finely tuned selective processing (through effective information integration and resource allocation) in the face of poor visual evidence.

Constraint satisfaction remains a major approach for bringing knowledge into real-time vision. In VIEWS, the main demonstration of behavioural evaluation and incident detection in traffic scenes used such techniques [37]. Furthermore, although knowledge-based vision has a poor history in robotics [11], innovative research by Mackworth [43] has shown that contraint-based vision can deliver a "quick and clean" response. In his situated agent approach, constraint nets specify robot behaviour in terms of both the goals and low-level reactions using a formal model that incorporates a symmetrical coupling of the robot with its environment. In situated cognition, the role of the environment is emphasised for active problem solving so that both the agent acting on the environment and the environment shaping the behaviour of the agent is fully modelled. Mackworth automatically constructs a constraint-satisfying controller from the formal model for the on-line system using a generalised dynamical system language. This use of more situated models, inspired by interdisciplinary research, is a promising, new direction in the subfield.

2.2 Model-based Reasoning

The model-based vision approach also has an early knowledge-based examplar, ACRONYM [10], which used symbolic reasoning to aid static scene interpretation. WALKER [29] was an early dynamic model-driven interpretation system that could identify examples of moving people in image sequences. In model-based vision, the stored knowledge is concerned with the expected objects, often specifying part-whole relationships and constraints among the subparts, but also relationships over time. The visual processing is driven by hypotheses, primarily top-down. For example, the ACRONYM system used stored models in the form of slot and filler frames which formed the nodes of the "object graph". Generalised cylinders were used as primitives in this hierarchical structure which represented objects from coarse to fine detail. Algebraic constraints could also be specified to build up the hierarchical "restriction graph". To drive the processing, ACRONYM constructed a "prediction graph" using these models and some reasoning. Then low-level edge and ribbon-like structures were constructed under the direction of the predictor module to form the "observation graph". Finally, the "interpretation graph" matched the observed features and relationships to the models using more reasoning to eliminate inconsistencies. Again, more recently, model-based vision systems have been refined using probabilistic techniques, for example [5].

Model-based vision techniques have also been refined by Koller and Nagel [39] using fully parameterised object models which can deliver detailed descriptions of tracked objects. Another important technique is to use 2D iconic representations from different views of the 3D model to simplify the matching. For example, Sullivan and colleagues [65, 70] have developed model-based tracking in traffic scenes for performance under real-time constraints. There is ongoing debate about the roles of iconic and 3D representations in the many different tasks performed by computer vision systems. Another notable development in model-based vision is the use of deformable objects which have to be described using statistical rather than geometric relationships [17, 64]. A major advantage of such representations is that they can be learnt from examples, as shown by the work of Baumberg and Hogg [3]. The use of iconic representations and statistical relationships, which can easily be acquired from images, is generally accepted to be biologically plausible. However, there are many open questions about the effectiveness of more formal analysis and the modelling of high-level invariance for computer vision tasks.

2.3 Logic Frameworks

In common with much work in AI, logic-based approaches have a great deal to offer in terms of consistency checking and explicit, declarative knowledge representation. In particular, formal approaches using well-defined languages with clear meaning for time, events, and causality, e.g., Allen [1] and Shoham [56], are useful for validating and prototyping new approaches in many AI subfields. For image interpretation, the reconstruction of MAPSEE within a logical framework [50] is a classic example. Spatial and temporal logics are characterised by declarative representation in some formal description language and reasoning using some form of theorem-proving or calculus. However, translating the knowledge into a precompiled procedural form for fast execution, as in the work of Kaelbling and Rosenschein [36], is a major trend in the field. For example, the work described earlier by Mackworth [43] using contraint-based vision in situated agents was based on underlying formal logic notions, so that the designer can achieve provably correct behaviour. The constraint net models were transformed into their dynamic forms to allow fast processing in the on-line system. This trend makes formal modelling extremely useful for robotics and in classifying objects and types of events, as well as for spatiotemporal reasoning in knowledge-based vision.

Recently, the need for formal descriptions in visual knowledge representation has been emphasised by Schroeder and Neumann [54]. They advocate the use of an object-centred, description logic tailored to the requirements of image understanding, together with an effective calculus. Their language can be used to formalise scene-independent domain knowledge using a set of axioms. However, there is still some way to go in making the calculus tractable for realistic problems. More applied work on spatiotemporal reasoning in VIEWS for advanced surveillance used logical rules which could be made into executable networks for incident detection [37] or used for occlusion reasoning [66]. Semantic regions underlying the interpretation of behavior in traffic scenes [22] and trajectories for event descriptions [35] have also been learnt from images to support high-level reasoning.

2.4 Probabilistic Frameworks

Also in common with more general AI, probabilistic approaches have much to offer in dealing with the pervasive problems of uncertainty and in allowing information integration. In probabilistic reasoning, the likelihood of classes of objects or events is inferred by propagation of belief values in the light of changing evidence. The two main frameworks for such reasoning used in knowledge-based vision are Bayesian Belief Networks (BBNs) advocated by Pearl [48] and the Dempster-Shafer theory of evidence [24]. For example, early work by Binford [5] used Bayesian inference to make model-based vision reliable while remaining computationally tractable. Dempster-Shafer theory has also been used [4], but the computational complexity of the scheme means that it is only practical at the level of conceptual evaluation. BBNs, on the other hand, have been more widely adopted in vision systems as they are applicable to all levels of the visual processing because of the fast updating possible with singly connected trees. For example, Rimey and Brown [52] used them to model geometric constraints for active control of camera movements and Gong and Buxton [27] grouped optic flow vectors for segmentation and tracking. BBNs provide a clear mapping of contextual knowledge onto the computation to constrain interpretation by combining known causal dependencies with estimated statistical knowledge. They also support closed-loop control and attentional processing using both top-down and bottom up messages in the propagation of belief values, as well as the possibility of learning and refining representations by observation [13].

Bayesian belief nets are now being used in many demanding applications such as BATmobile [23] and TEA [53] to provide essential information integration. Buxton and Gong [13] have also developed a systematic methodology for the design, integration and implementation of advanced vision systems using BBNs. These networks allow dynamic updating of values in visual evidence and interpretation nodes, but not specification of the temporal constraints themselves. Howarth and Buxton [32] used dynamically reconfigured networks to model the evolving spatial relationships of vehicles as they move through the scene. Others [23] have adopted the dynamic probabilistic networks developed by Dean and Kanazawa [21] which make use of the simple Markov property that the future is independent of the past given the present. BBNs support the active control of visual processing and off-line learning of the prior and conditional probabilities in many applications, see Spiegelhalter and Cowell [58]. They have even been learnt on-line using reinforcement learning by Whitehead and Ballard [69]. These approaches, then, are very promising for advanced vision systems that require ongoing exploitation and acquisition of knowledge.

3 Exploitation

As discussed in the introduction, the task-specific requirements of applied vision systems often drive the development of high-level vision capabilities. Thus, a great deal of innovative research in interpretation and understanding is both developed and exploited in a variety of application contexts. For example, there has been important research to integrate vision and language and deliver conceptual descriptions for advanced surveillance. Pioneering research on describing behaviour in traffic scenes by Nagel [46] and Neumann [47] established a useful ontology for the events and episodes observed. More recently, this has been extended in terms of both the complexity of vehicle interactions analysed by Howarth and Buxton [14, 33] and the sophistication of the linguistic descriptions computed by Nagel and colleagues [25, 40]. Real-time constraints for descriptions in video-surveillance applications have also received attention in the new PASSWORDS project [16]. These techniques were clearly developed for advanced surveillance but are also more generally applicable in interactive vision systems.

Suchman [63] proposed a situated approach for general human computer interaction and here, again, there is a clear requirement for systems that integrate both vision and language, for example [55]. Interdisciplinary work in cognitive science, HCI, and AI approaches to vision and language will be an important component of long term work in this area. In the short term, many researchers are developing useful techniques for multimodal and multimedia interaction. For example, Kender [38] has been active in bringing spatial reasoning and gesture recognition to these problems. Bobick [8, 9] has also been leading work at MIT Media lab for a variety of interactive vision applications. These applications seek to understand actions directly from the image sequences using approximate models in order to meet real-time constraints.

Smart cars using new sensor technology for vehicle control are also being developed in conjuction with traffic monitoring in intelligent highway system projects by Malik and colleagues [34, 44]. This type of application is also closely linked to innovative work on behavioural control in robotics by Bajcsy and colleagues [41, 57] using discrete event dynamic systems. The idea of integrating work on understanding scenes with behavioural control for automatic vehicle guidance has great commercial potential and exciting new work is being done in this area. In addition, a new situated approach using constraint-based vision by Mackworth [43] is being developed to integrate knowledge-based and behavioural control in robotics. These developments, then, involve fundamental science while being highly applicable for real-world applications.

4 New Directions

In conclusion, there are many new directions that seem promising for this rapidly expanding subfield. We have seen that probabilistic reasoning is being used to provide effective integration, allowing representation of context, control, and even learning. The use of iconic representations, which are easily acquired from example images and can be used in subsequent recognition of the objects and behaviours, is also a major new direction. In a very different direction, there is a requirement to formalise reasoning to provide provably correct behaviour in many applications. This requires close interaction of specialised subfields using logic in AI and high-level vision research. Another move in this direction is the integration of work in vision and language for many application areas in advanced surveillance, medical analysis systems, and multimodal interaction. Work on such interactive systems forces the developers to use frameworks with a common semantics and to adopt cognitive models of the system users.

In addition to the directions above, there are new themes that are beginning to influence work on knowledge-based vision. In particular, the combination of deformable models with dynamic learning of their statistical properties seems set to grow rapidly. As we have seen, there are many applications of deformable models in biomedical image analysis [17], face recognition [42], and tracking of people [3]. The acquisition of these models by training can allow the development of generative models and these are now starting to model physical forces for visual understanding. In addition, situated cognition is now being taken seriously in the interpretation and understanding community, although there is a rejection of strong anti-representational positions like that of Brooks [11]. For example, work by Howarth and Buxton [13, 30] on situated behavioural analysis and work by Mackworth [43] on situated agents for robotics. These examples are just a part of the groundshift over the last two decades from the traditional approach based on symbolic reasoning in Good Old-Fashioned Artificial Intelligence (GOFAI) to simpler, behaviour-based approaches.

References

[1] J.F. Allen. "Towards a general theory of action and time". *Artificial Intelligence*, 23:123–154, 1984.

[2] D.H. Ballard. "Animate vision". *Artificial Intelligence*, 48:57–86, 1991.

[3] A. Baumberg and D.C. Hogg. "Generating spatiotemporal models from examples". In *British Machine Vision Conference*, Birmingham, UK, 1995.

[4] B. Besserer, S. Estable, and B. Ulmer. "Multiple knowledge sources and evidential reasoning for shape recognition". In *International Conference on Computer Vision*, Berlin, Germany, 1993.

[5] T.O. Binford, T.S. Levitt, and W.B. Mann. "Bayesian inference in model-based machine vision". In *Uncertainty in Artificial Intelligence 3*. Machine Intelligence and Pattern Recognition Series Volume 8, North-Holland, 1989.

[6] L. Birnbaum, M. Brand, and P. Cooper. "Looking for trouble: Using causal semantics". In *International Conference on Computer Vision*, Berlin, Germany, 1993.

[7] A. Blake, M. Isard, and D. Reynard. "Learning to track the visual motion of contours". *Artificial Intelligence*, 78:179–212, 1995.

[8] A.F. Bobick and C. Pinhanez. "Using approximate models as a source of contextual information for vision processing". In *Workshop on Context-based Vision*. IEEE Press, 1995.

[9] A.F. Bobick. "Computers seeing action". In *British Machine Vision Conference*, Edinburgh, Scotland, 1996.

[10] R.A. Brooks. "Symbolic reasoning among 3D models and 2D images". *Artificial Intelligence*, 17:285–348, 1981.

[11] R.A. Brooks. "Elephants don't play chess". *Robotics and Autonomous Systems*, 6:3–15, 1990.

[12] H. Buxton et al. "VIEWS: Visual Inspection and Evaluation of Wide-area Scenes". In *12th IJCAI Videotape Program*. Morgan Kaufmann, 1991.

[13] H. Buxton and S. Gong. "Visual surveillance in a dynamic and uncertain world". *Artificial Intelligence*, 78:371–405, 1995.

[14] H. Buxton and R. Howarth. "Spatial and temporal reasoning in generation of dynamic scene descriptions". In *IJCAI Workshop on Spatial and Temporal Reasoning*, Montreal, Canada, 1995.

[15] H. Buxton and R. Howarth. "Watching behaviour: The role of context and learning". In *International Conference on Image Processing*, Lausanne, Switzerland, 1996.

[16] N. Chleq and M. Thonnat. "Realtime image sequence interpretation for video surveillance applications". In *International Conference on Image Processing*, Lausanne, Switzerland, 1996.

[17] T.J. Cootes, C.J. Taylor, D.H. Cooper, and J. Graham. "Training models of shape from sets of examples". In *British Machine Vision Conference*, Leeds, UK, 1992.

[18] D.R. Corrall and A.H. Hill. "Visual surveillance". *GEC Review*, 8:15–27, 1992.

[19] J. Crowley, J.M. Bedrune, M. Bekker, and M. Schneider. "Integration and control of reactive visual processes". In *European Conference on Computer Vision*, Stockholm, Sweden, 1994.

[20] J.L. Crowley and H. Christensen. *Vision as Process.* Springer-Verlag, Berlin, 1993.

[21] T. Dean and K. Kanazawa. "Probabilistic temporal reasoning". In *National Conference on Artificial Intelligence*, AAAI Press, 1988.

[22] J.H. Fernyhough, A.G. Cohn, and D.C. Hogg. "Generation of semantic regions from image sequences". In *European Conference on Computer Vision*, Cambridge, UK, 1996.

[23] J. Forbes, T. Huang, K. Kanazawa, and S. Russell. "The BATmobile: Towards a Bayesian automated taxi". In *International Joint Conference on Artificial Intelligence*, pages 1878–1885, Montreal, Canada, 1995.

[24] T.D. Garvey, J.D. Lowrance, and M.A. Fischler. "An intelligence technique for integrating knowledge from disparate sources". In *International Joint Conference on Artificial Intelligence*, Vancouver, Canada, 1981.

[25] R. Gerber and H.H. Nagel. "Knowledge representation for the generation of quantified natural language descriptions of vehicle traffic in image sequences". In *International Conference on Image Processing*, Lausanne, Switzerland, 1996.

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

[27] S. Gong and H. Buxton. "Bayesian nets for mapping contextual knowledge to computational constraints in motion segmentation and tracking". In *British Machine Vision Conference*, Guildford, UK, 1993.

[28] A.R. Hanson and E.M. Riseman. *Computer Vision Systems.* Academic Press, New York, 1978.

[29] D.C. Hogg. "Model-based vision: A program to see a walking person". *Image and Vision Computing*, 1:5–21, 1983.

[30] R.J. Howarth. "Interpreting a dynamic and uncertain world: high-level vision". *Artificial Intelligence Review*, 9:37–63, 1995.

[31] R.J. Howarth and H. Buxton. "An analogical representation of space and time". *Image and Vision Computing*, 10:467–478, 1992.

[32] R.J. Howarth and H. Buxton. "Selective attention in dynamic vision". In *International Joint Conference on Artificial Intelligence*, Chambery, France, 1993.

[33] R.J. Howarth and H. Buxton. "Visual surveillance monitoring and watching". In *Computer Vision -ECCV'96.* Springer Verlag, 1996.

[34] T. Huang, D. Koller, J. Malik, G. Ogasawara, S. Russell and J.Weber. "Automatic symbolic traffic scene analysis using belief networks". In *National Conference on Artificial Intelligence*, AAAI Press, 1994.

[35] N. Johnson and D.C. Hogg. "Learning the distribution of object trajectories for event recognition". In *British Machine Vision Conference*, Birmingham, UK, 1995.

[36] L.P. Kaelbling and S.J. Rosenschein. "Action and planning in embedded agents". *Robotics and Autonomous Systems*, 6:35–48, 1990.

[37] S. King, S. Motet, J. Thoméré, and F. Arlabosse. "A visual surveillance system for incident detection". In *AAAI Workshop on AI in Intelligent Vehicle Highway Systems*, AAAI Press, 1994.

[38] R. Kjeldsen and J. Kender. "Knowledge-based hand gesture recognition". In *Workshop on Context-based Vision.* IEEE Press, 1995.

[39] D. Koller, K. Danilidis and H.H. Nagel. "Model-based object tracking in monocular image sequences of road traffic". *International Journal of Computer Vision*, 10:257-281, 1993.

[40] H. Kollnig, M. Otte, and H.H. Nagel. "Association of motion verbs with vehicle movements extracted from dense optical flow fields". In *European Conference on Computer Vision*, Stockholm, Sweden, 1994.

[41] J. Kosecka and R. Bajcsy. "Cooperation of visually guided behaviours". In *International Conference on Computer Vision*, Berlin, Germany, 1993.

[42] A. Lanitis, C.J. Taylor, and T.F. Cootes. "A unified approach to coding and interpreting face images". In *International Conference on Computer Vision*, Cambridge, MA, 1995.

[43] A. Mackworth. "Quick and clean: Constraint-based vision for situated robots". In *International Conference on Image Processing*, Lausanne, Switzerland, 1996.

[44] J. Malik, J. Weber, O.T. Luong and D. Koller. "Smart cars and smart roads". In *British Machine Vision Conference*, Birmingham, UK, 1995.

[45] M. Mohnhaupt and B. Neumann. "Understanding object motion: Recognition, learning and spatiotemporal reasoning". *Journal of Robotics and Autonomous Systems*, 8:65–91, 1991.

[46] H.H. Nagel. "From image sequences towards conceptual descriptions". *Image and Vision Computing*, 6:59–74, 1988.

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

[48] J. Pearl. "Distributed revision of composite beliefs". *Artificial Intelligence*, 33:173–215, 1987.

[49] R.P.N. Rao and D.H. Ballard. "An active vision architecture based on iconic representations". *Artificial Intelligence*, 78:461–506, 1995.

[50] R. Reiter and A.K. Mackworth. "A logical framework for depiction and image interpretation". *Artificial Intelligence*, 41:125–155, 1989.

[51] R.D. Rimey and C.M. Brown. "Selective attention as sequential behavior: Modeling eye movements with an augmented Hidden Markov Model". *Computer Science TR327, University of Rochester*, 1990.

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

[53] R.D. Rimey and C.M. Brown. "Control of selective perception using Bayes nets and decision theory". *International Journal of Computer Vision*, 12:173–209, 1994.

[54] C. Schroeder and B. Neumann. "On the logics of image interpretation: Model construction in a formal knowledge representation framework". In *International Conference on Image Processing*, Lausanne, Switzerland, 1996.

[55] G. Socher, G. Sagerer, F. Kummert and T. Fuhr. "Talking about 3D scenes: Integration of image and speech understanding in a hybrid distributed system". In *International Conference on Image Processing*, Lausanne, Switzerland, 1996.

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

[57] T.M. Sobh and R. Bajcsy. "Visual observation as a discrete event dynamic system". In *IJCAI Workshop on Dynamic Scene Understanding*, Sydney, Australia, 1991.

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

[59] L. Stark and K. Bowyer. "Functional context in vision". In *Workshop on Context-based Vision.* IEEE Press, 1995.

[60] T.M. Strat and M.A. Fischler. "Context-based vision: Recognising objects using both 2D and 3D imagery". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13:1050–1065, 1991.

[61] T.M. Strat and M.A. Fischler. "The role of context in computer vision". In *Workshop on Context-based Vision.* IEEE Press, 1995.

[62] R.K. Srihari. "Linguistic context in vision". In *Workshop on Context-based Vision.* IEEE Press, 1995.

[63] L. Suchman. *Plans and Situated Action: The Problems of Human-Machine Communication.* Cambridge University Press, Cambridge, 1987.

[64] G.D. Sullivan, A. Worrall, and J.M. Ferryman. "Visual object recognition using deformable models of vehicles". In *Workshop on Context-based Vision*. IEEE Press, 1995.

[65] G.D. Sullivan, K.D. Baker, A. Worrall, C.I. Attwood, and P.R. Remagnino. "Model-based vehicle detection and classification using orthographic approximations". In *British Machine Vision Conference*, Edinburgh, Scotland, 1996.

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

[67] J.K. Tsotsos. "Knowledge organisation and its role in representation and interpretation for time-varying data: The ALVEN system". *Computing Intelligence*, 1:498–514, 1985.

[68] J.K. Tsotsos, S.M. Culhane, W.Y.K. Wai, Y. Lai, N. Davies, and F. Nuflo. "Modeling visual attention via selective tuning". *Artificial Intelligence*, 78:507–545, 1995.

[69] S.D. Whitehead and D.H. Ballard. "Learning to perceive and act by trial and error". *Machine Learning*, 7:45–83, 1991.

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