Visual Interpretation and Understanding
CSRP 452

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

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 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 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 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 are

[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".

[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 gen- temporal
tional

[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 [illegible]

[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

[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.