Task and Behaviour Learning for ActIPret Project

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

December 21, 2001

Abstract

In this position paper we propose behaviour learning and recognition techniques to be used in advanced vision applications. Bayesian Belief Networks (Bayes nets) have been developed for cognitive vision and support information integration and representation of contextual constraints. They also work well with the task-based control proposed for ActIPret using DDNs. Radial Basis Function Networks (RBF nets) have also been developed for more reactive vision tasks and work well for fast learning and classification. We propose a combination of these two approaches for event, action and activity learning and representation. In addition, specific extensions of our existing work to allow more general 3D activity analysis in the 'attentive' processing are: 1) 'modular' Bayes nets for multi-agent learning and recognition ('agent' is movable person/object); 2) development of the new interface with task control policies (specifically the evaluation criteria); 3) action-based representation in a hand frame-of-reference; and 4) more general 'parsing' of activity plans in the later stages of the project. For 'pre-attentive' processing, proposed extension is to adapt the time-delay RBF net scheme to use 3D trajectory information in gesture recognition and other primitive visual operators used to cue full attentional processing.

1 Introduction

Research issues to be addressed in building cognitive vision systems centre on the role of context, control and learning. Four major approaches are 1) logic-based approaches, 2) graph-based models, 3) deformable models and 4) neural networks. Of these, it is primarily approaches 2-4 that have associated learning theory, usually based on statistical techniques of some kind. By now it is clear that in ActIPret, it is best to have learnt representations so that our systems can scale up. It has also been proposed that some global, less controlled processing (pre-attentive) establishes a set of candidate objects and predicitive action cues

in the system before committing resources to full 'attentional' processing in the online system. Finally, it has been proposed that ActIPret requires primarily 3D vision. Thus, this position paper first introduces general visual learning, then proposals for task and behaviour learning at the 'attentive' and 'pre-attentive' levels. It also discusses extensions to existing work to deal with the specific requirements of the ActIPret system.

Learning in a vision system can be at the level of object models, their movements and actions, and how to control views and processing in the system. Our work on appearance-based approaches (using RBF nets) suggests they are more learnable and robust than structural approaches for general object categorisation on real-world tasks such as face recognition [23, 25]. Natural deformable objects are difficult to specify and so are their movements and actions, so adaptive methods are required. At the heart of a visual learning system is the ability to find the relevant mapping from observable or derivable attributes of image(s) onto the visual categories we require for real-world tasks. In the proposal below, we show how appearance-based techniques can be extended to gesture recognition in ActIPret.

In addition to object and behaviour recognition, expected behaviour can be used to control further processing in the system through prediction. Many different learning and prediction techniques have been proposed. For example, symbolic learning using case-based reasoning [12], graphical models for probabilistic reasoning and control [18, 38], stochastic models for learning and prediction in tracking [5], deformable models for event analysis [2, 31], neural network learning in gesture recognition [24]. We will examine these recent developments in more detail in the next section and propose future extensions for ActIPret.

2 Proposal

2.1 Cognitive recognition and synthesis

Probabilistic frameworks have much to offer in dealing with the pervasive problem of uncertainty in visual evidence, and support information integration and learning as discussed by Pearl [32]. The representation of constraints in a Bayesian Belief Network, 'Bayes net', can be achieved by mapping into the structure of a Directed Acyclic Graph (DAG) so that the nodes represent concepts of interest and dependencies are given by causal links. A simpler chained structure with single causal dependencies over time, the Hidden Markov Model (HMM), is often used for speech analysis [36] and has been extensively adapted for analysis of dynamic scenes and perceptual control as described below. In probabilistic reasoning [9], the likelihood of classes of objects or events is inferred by propagation of belief values in the light of changing evidence. Dempster-Shafer theory has also been used to handle uncertainty in vision [3] but this approach is generally more computationally complex in the online evaluation.

Bayes nets have been widely adopted in vision systems as they are applicable to all levels of processing due to fast numerical updating in singly connected

trees. There are techniques to decompose complex models and handle networks with multiple causes as well as learn the parameters for such networks [40]. Rimey and Brown introduced such techniques for active vision, with both control of camera movement and the selective processing required in task-based perceptual control [37, 38]. As discussed by Gong and Buxton [18], Bayes nets provide a clear way to map contextual constraints from the scene onto the computation of the visual interpretation by combining known causal dependencies with estimated statistical knowledge. They are essentially providing closed-loop control using both top-down and bottom-up messages in the propagation of belief values. They also provide the possibility of learning and refining visual representations by observation [8, 41]. Bayes nets have been used in many demanding applications such as BATmobile [13] and TEA system [38].

HMMs are also widely used in visual processing. The advantage here is that the "hidden" purposes of regular behaviour patterns can be learnt from examples, i.e. the structure of the model as well as the parameters are easily learnt [36]. For example, in early work by Gong [17, 16] the movement patterns of vehicles on an airport ground-plane were learnt to provide a model used in prediction. More recent work uses parameterised HMMs for gesture interpretation [6], coupled HMMs to learn models for interactions [7], and variable length HMMs for virtual reality systems [14].

For the cognitive level behaviour integration then, it seems that HMMs and Bayes nets have much to offer in the 'attentional' processing. We can extend existing work by:

- First, exploiting ideas from Nevatia's group who have used a Bayesian approach to develop efficient, modular multi-agent event recognition [19, 20]. However, we would not use agent- based 'threads' as they suggest but rather retain our 'situated vision' visual index approach [21, 22]. This involves close interaction with deictic relationships derived in WP3.

- Second, the modular approach to event/activity recognition will require extensions to the interface with task-based control in the evaluation loop that selects and schedules visual operations and confirms/rejects the hypotheses concerning current behaviour in WP1.

- Third, as I emphasised in the kick-off meeting, the approach to behaviour modelling needs to be generalised for action-based representations. While traffic analysis can take place on a ground-plane, 3D activities in ActIPret require representation and reasoning in the hand frame-of-reference.

- Fourth, when there are multiple, valid Activity Plans, a more general reactive planning or 'parsing' mechanism will be required in the task-control policies during behaviour recognition.

2.2 Detection and early behaviour cues

Neural network techniques are a powerful, general approach to pattern recognition tasks and there are a variety of different methods (for an introduction see

[4]). The classical networks do not include a time dimension so they have to be adapted to deal with dynamic scene analysis. Some extended models have internal time like the partially recurrent networks of Elman [11] and Jordan [28]. Others have external time like the time-delay networks described below. Time can be explicitly represented in the architecture at the network level using the connections or can be represented at the neuron level, including the recently developed 'spiking networks'. These model the intrinsic temporal properties of biological neurons, which fire with a pattern of pulses or spikes. The most common new generation dynamic network of this kind is the 'integrate-and-fire' (IF) network (for review see [15]). However, they have yet to be applied in visual behaviour analysis as there is still ongoing debate about how best to propagate information (the 'coding problem') in these models. An important exception is the extension based on classical Radial Basis Functions (RBFs).

The RBF net is a two-layer, hybrid learning network [29, 30], which combines a supervised layer from the hidden to the output units with an unsupervised layer from the input to the hidden units. The network model is characterised by individual radial Gaussian functions for each hidden unit, which simulate the effect of overlapping and locally tuned receptive fields. It is characterised by computational simplicity, supported by well-developed mathematical theory, and robust generalisation, powerful enough for real-time, real-life tasks [35, 39]. The nonlinear decision boundaries of RBF nets make better general function approximations than the hyperplanes created by the multi-layer perceptron (MLP) with sigmoid units [33], and they provide a guaranteed, globally optimal solution via simple, linear optimisation. One advantage of the RBF net, compared to the MLP, is that it gives low false-positive rates in classification problems as it will not extrapolate beyond its learnt example set. This is because its basis functions cover only small localised regions, unlike sigmoidal basis functions which are nonzero over an arbitrarily large region of the input space. RBF nets are also fairly robust to partial occlusions [1].

Once training examples have been collected as input-output pairs, with the target class attached to each image, tasks can be learnt directly by the system. This type of supervised learning can be seen in mathematical terms as approximating a multivariate function, so that estimations of function values can be made for previously unseen test data where actual values are not known. This process can be undertaken by the RBF net using a linear combination of basis functions, one for every training example, because of the smoothness of the manifold formed by the example views of objects in a space of all possible views of that object [34]. This underlies successful previous work with RBF nets for face recognition from video sequences [25], which uses an RBF unit for each training example, and rapid pseudo-inverse calculation of weights. An important factor in this approach is the flexibility of the RBF net learning approach, which allows formulation of the training in terms of the specific classes of data to be distinguished. For example, extraction of identity, head pose and expression information can be performed separately on the same face training data to learn a computationally cheap RBF classifier for each separate recognition task [10, 26]. Essentially, these adaptive methods allow key inferences to be made

within the system by modeling the variability of the evidence.

To extend this research to support *Visually Mediated Interaction* (VMI), person-specific and generic gesture models were developed for the control of active cameras. A time-delay variant of the Radial Basis Function (TDRBF) net recognised pointing and waving hand gestures in image sequences [24, 27]. A gesture database was developed as a source of suitable image sequences for these experiments. Characteristic visual evidence is automatically selected during the adaptive learning phase, depending on the task demands. A set of interaction-relevant gestures were modeled and exploited for reactive on-line visual control. These were then interpreted as user intentions for live control of an active camera with adaptive view direction and attentional focus. For ActIPret, some of the ideas for zooming in on activities can still be exploited. Also the gesture recognition is an excellent predictive cue for many of the actions and activities in our ActIPret scenarios. At the earlier levels of processing, but particularly in the gesture recognition, reactive behaviour is important for both camera movement and invoking further 'attentional' processing. The current scheme is entirely 'appearance-based' using RBF unit 'prototypes' that deliver a confidence measure of how likely new data fits this unit's learnt function. These first layer outputs are combined in task-specific ways to deliver classes or drive camera views. Thus, specific extensions for ActIPret are: 1) adapt the TDRBF net scheme to accept 3D hand trajectories for predictive gesture recognition (or possibly a multi-view appearance-based scheme but this requires training data from a set of viewpoints). The gesture recognition can use pre- and mid-gesture phase detectors as in our previous work on predictive control and requires interfaces to the visual index mechanism of WP1 to know which trajectory corresponds to the hand (WP4 delivers this); 2) extend the gesture scheme for two hands under WP5 (although the multi-hypotheses are likely to be handled better by 'attentive' Bayes net); and 3) with partners decide at which level to drive 'attentional' camera movements (WP6) as it may be appropriate to drive this as an early reactive behaviour.

3 Conclusion

I hope this position paper clarifies aspects of the approach, but not the detail, for the Task and Behaviour WP5 learning and recognition in ActIPret. Please comment urgently as it is crucial for the planned work of Jon Howell in the COGS team. We have a great deal of work to do here and need to have a working framework with at least primitive gesture recognition and isolated activity recognition in place by end of year 1. We propose to schedule work on TDRBF gesture recognition as soon as 3D hand trajectories are available from FORTH (interaction with WP4) and agree an initial position on camera control with PROFACTOR (interaction with WP6) asap. We also propose to start immediately on the work on modular Bayes nets for simple event and activity recognition (interaction with deictic relationships WP3). This will then allow us to confirm completed actions for synthesis of the Activity Plan in the expert

mode.

The work on extensions to the interface with WP1 in the task-control policies in evaluation confirming or rejecting behaviour hypotheses must also exist in an initial version by the end of year 1. Also, the representation scheme for the hand-based activities is fundamental to learning and recognising activities so must be taken as a starting point for WP5 work at the cognitive levels. Finally, later in year 2/3 we can extend this work to allow a more complete set of valid Activity Plans in a flexible, learning and interpretation scheme. This will involve multiple hypotheses in some kind of Bayesian framework, or possibly Finite State (FS) machine, to 'parse' the ongoing interactions in the dynamic scene, similar to natural language analysis. Also, in year 3, we can extend the gesture and primitive operator set for a wider class of predictive cues using 2 hands.

I have suggested extensions required in WP5 and their interactions with other WPs, especially Cognitive Framework WP1 and Deictic Spatial and Temporal Relationships WP3. In addition, we still have much work in thinking out the interaction with Attentive and Investigative Behaviours WP6. We also note that WP3 itself interacts strongly with WP2 and WP4 on object recognition and tracking at the cognitive levels to derive the relevant spatial and temporal relationships for behaviour analysis on which WP5 depends.

References

[1] S. Ahmad and V. Tresp. "Some solutions to the missing feature problem in vision". In S. J. Hanson, J. D. Cowan, and C. L. Giles, editors, *Advances in Neural Information Processing Systems*, volume 5, pages 393–400, 1993.

[2] A. Baumberg and D.C. Hogg. "Generating spatiotemporal models from training examples". *Image and Vision Computing*, 14:525–532, 1996.

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

[4] C.M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.

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

[6] A. Bobick and A. Wilson. "A state-based technique for the summarization and recognition of gesture". In *IEEE International Conference on Computer Vision*, pages 382–388, Cambridge, MA, 1995.

[7] M. Brand, N. Oliver, and A. Pentland. "Coupled hidden Markov models for complex action recognition". In *IEEE Conference on Computer Vision and Pattern Recognition*, San Juan, Puerto Rico, 1997.

[8] H. Buxton and S. Gong. "Visual surveillance in a dynamic and uncertain world". *Artificial Intelligence*, 78:431–459, 1995.

[9] E. Charniak. "Bayesian networks without tears". *AI Magazine*, 12(4):50–63, 1991.

[10] S. Duvdevani-Bar, S. Edelman, A. J. Howell, and H. Buxton. "A similarity-based method for the generalization of face recognition over pose and expression". In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 118–123, Nara, Japan, 1998.

[11] J. Elman. "Finding structure in time". *Cognitive Science*, 14:179–211, 1990.

[12] J.H. Fernyhough, A.G. Cohn, and D.C. Hogg. "Constructing qualitative event models automatically from video input". *Image and Vision Computing*, 18:81–103, 2000.

[13] 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.

[14] A. Galata, N. Johnson, and D.C. Hogg. "Learning variable length Markov models of behaviour". *Computer Vision and Image Understanding*, 81:398–413, 2001.

[15] W. Gerstner. "Time structure of the activity in neural networks". *Physical Review*, E 51:738–758, 1995.

[16] S. Gong. "Visual behaviour: Modelling "hidden" purposes in motion". In *SPIE International Conference on Neural and Stochastic Methods in Image and Signal Processing*, San Diego, CA, 1992.

[17] S. Gong. "Visual observation as reactive learning". In *SPIE International Conference on Adaptive and Learning Systems*, pages 175–187, Orlando, FL, 1992.

[18] S. Gong and H. Buxton. "Bayesian nets for mapping contextual knowledge to computational constraints in motion segmentation and tracking". In *British Machine Vision Conference*, pages 229–238, Guildford, UK, 1993.

[19] S. Hongeng, F. Bremond, and R. Nevatia. "Bayesian framework for video surveillance applications". In *International Conference on Pattern Recognition*, Barcelona, Spain, 2000.

[20] S. Hongeng and R. Nevatia. "Multi-agent event recognition". In *IEEE International Conference on Computer Vision*, Vancouver, Canada, 2001.

[21] R. Howarth and H. Buxton. "Visual surveillance monitoring and watching". In *European Conference on Computer Vision*, pages 321–334, Cambridge, UK, 1996.

[22] R. Howarth and H. Buxton. "Conceptual descriptions from monitoring and watching image sequences". *Image and Vision Computing*, 18:105–135, 2000.

[23] A.J. Howell and H. Buxton. "Invariance in Radial Basis Function networks in human face classification". *Neural Processing Letters*, 2:26–30, 1995.

[24] A.J. Howell and H. Buxton. "Learning gestures for visually mediated interaction". In *British Machine Vision Conference*, Southampton, UK, 1998.

[25] A.J. Howell and H. Buxton. "Learning identity with Radial Basis Function networks". *Neurocomputing*, 20:15–34, 1998.

[26] A.J. Howell and H. Buxton. "Time-delay RBF networks for attentional frames in Visually Mediated Interaction". *Neural Processing Letters*, 2001.

[27] A.J. Howell and H. Buxton. "Active vision techniques for Visually Mediated Interaction". *Image and Vision Computing*, 2002.

[28] M. I. Jordan. "Serial order: A Parallel, Distributed Processing approach". In J. L. Elman and D. E. Rumelhart, editors, *Advances in Connectionist Theory: Speech*. Lawrence Erlbaum, 1989.

[29] J. Moody and C. Darken. "Learning with localized receptive fields". In D. Touretzky, G. Hinton, and T. Sejnowski, editors, *Proceedings of 1988 Connectionist Models Summer School*, pages 133–143, 1988.

[30] J. Moody and C. Darken. "Fast learning in networks of locally tuned processing units". *Neural Computation*, 1:281–294, 1989.

[31] R.J. Morris and D.C. Hogg. "Statistical models of object interaction". *International Journal of Computer Vision*, 37:209–215, 2000.

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

[33] T. Poggio and S. Edelman. "A network that learns to recognise three-dimensional objects". *Nature*, 343:263–266, 1990.

[34] T. Poggio and F. Girosi. "Regularisation algorithms for learning that are equivalent to multilayer networks". *Science*, 247:978–982, 1990.

[35] D. A. Pomerleau. "ALVINN: An Autonomous Land Vehicle in a Neural Network". In D. S. Touretzky, editor, *Advances in Neural Information Processing Systems*, volume 1, pages 305–313, 1989.

[36] L.R. Rabiner. "A tutorial on hidden Markov models". *Proceedings of the IEEE*, 77:257–286, 1989.

[37] R.D. Rimey and C.M. Brown. "Where to look next using a Bayes net: Incorporating geometric relations". In *European Conference on Computer Vision*, pages 542–550, Genoa, Italy, 1992.

[38] 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.

[39] M. Rosenblum, Y. Yacoob, and L.D. Davis. "Human emotion recognition from motion using a radial basis function network architecture". *IEEE Transactions on Neural Networks*, 7:1121–1138, 1996.

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

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