Active Vision Techniques for Visually Mediated Interaction

A. Jonathan Howell and Hilary Buxton

CSRP 541

November 2001

ISSN 1350–3162

UNIVERSITY OF

Cognitive Science
Research Papers

Active Vision Techniques for Visuall Mediated Interaction

A. Jonathan Howell and Hilary Buxton

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

Abstract

In this paper we introduce adaptive vision techniques used, for example, in video-conferencing applications. First, we present the recognition of identity, expression and head pose using Radial Basis Function (RBF) networks. Second, we address gesture-based communication and attentional focus, using colour/motion cues to direct face detection and capture 'attentional frames'. These focus the processing for Visually Mediated Interaction via an appearance-based approach with Gabor filter coefficients used as input to time-delay RBF networks. Third, we present methods for the gesture recognition and behaviour (user-camera) coordination in an integrated system.

Key ords: Visually Mediated Interaction; Face Recognition; Gesture Recognition; Camera Control; Time-Delay Neural Networks

1 Introduction

Visually Mediated Interaction (VMI) is a process of facilitating interaction between people, either remotely or locally, using visual cues for discourse/interaction management. In particular, gaze direction is often associated with diectic, attention-directing pointing to indicate objects or people of interest in the immediate context as part of the behavioural interaction.

We know that robust tracking of non-rigid objects such as human faces and bodies involved in machine analysis of this kind of interactive activity is difficult due to rapid motion, occlusion and ambiguities in segmentation and model selection. This was partially addressed by the move to active vision and dynamic models for robust tracking using sophisticated Kalman filters, as exemplified by Blake and others [1]. Recently, these have been specialised to allow the learning of complex hand dynamics [23]. More generally, research funded by British Telecom (BT) on *Smart Rooms* [38] and the ALIVE project [30] at MIT Media Lab has shown progress in the modelling and interpretation of human body activity. This used the *Pfinder* (Person Finder) system [49], which can provide real-time human body analysis. Further analysis to model the progression of ongoing activity involves techniques such as *Hidden Markov Models*

2 The RBF Network Scheme

The RBF network is a two-layer, hybrid learning network [32,33], 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.

The RBF network is characterised by computational simplicity, supported by well-developed mathematical theory, and robust generalisation, powerful enough for real-time real-life tasks [42,43]. The nonlinear decision boundaries of the RBF network make it better in general for function approximation than the hyperplanes created by the multi-layer perceptron (MLP) with sigmoid units [41], and they provide a guaranteed, globally optimal solution via simple, linear optimisation. One advantage of the RBF network, compared to the MLP, is that it gives low f44Td(linear)Tj8-ed

Table 1
Body movement and behaviour definitions for the gesture database.

Gesture	Body Movement	Behaviour
pntrl	point right hand to left	pointing left
pntrr	point right hand to right	pointing right
wavea	wave right hand above head	urgent wave
waveb	wave right hand below head	non-urgent wave

Previous approaches to recognising human gestures from real-time video as a nonverbal modality for human-computer interaction have involved computing low-level features from motion to form *temporal trajectories* that can be tracked by Hidden Markov Models or Dynamic Time Warping. However, for this work we explored the potential of using simple image-based differences from video sequences in conjunction with the RBF network learning paradigm to account for variability in the appearance of a set of predefined gestures. The computational simplicity and robust generalisation of our alternative RBF approach provides giv0Td(p34ional)Tj76.f12.60T30.12Tf54.71990Td(geTd(the)Tj14.71990Td98(1jsui-

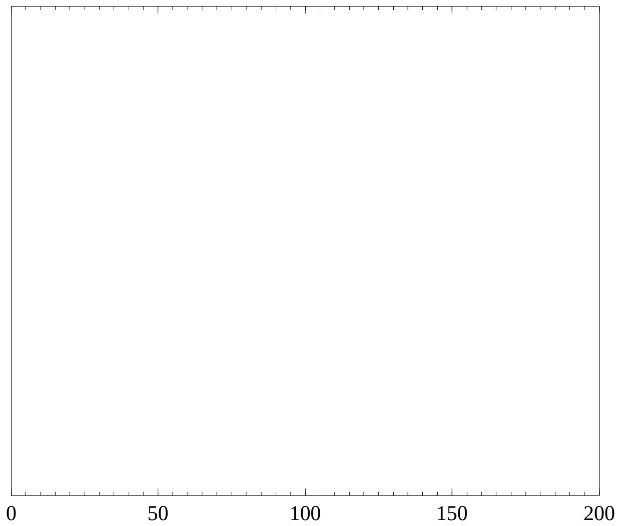

0
50
100
150
200

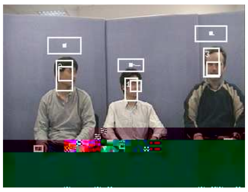
Table 2
Example interpretations of camera position vectors for group interaction scenarios with three people.

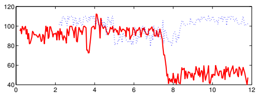
Camera Position Vector	Interpretation
`[0,0,0]`	frame whole scene
`[1,0,0]`	focus on subject A
`[0,1,1]`	focus on subjects B and C using a split-screen effect

While full computer understanding of dynamic visual scenes containing several people may be currently unattainable, we have investigated a computationally efficient approach to determine areas of interest in such scenes. Specifically, we have devised a method for modelling and interpretation of single- and multi-person human behaviour in real time to control video cameras [44]. Such machine understanding of human motion and behaviour is currently a key research area in computer vision, and has many real-world applications. *Visually Mediated Interaction* (VMI) is particularly important to applications in video telecommunications. VMI requires intelligent interpretation of a dynamic visual scene to determine areas of interest for effective communication to remote users.

As we have seen, our general approach to modelling behaviour is *appearance-based* in order to provide real-time behaviour interpretation and prediction [20,44]. In addition, we only use defined as any body movement sequence that is performed subconsciously by the participant, and here, it is head pose that is the primary source of implicit behaviour.

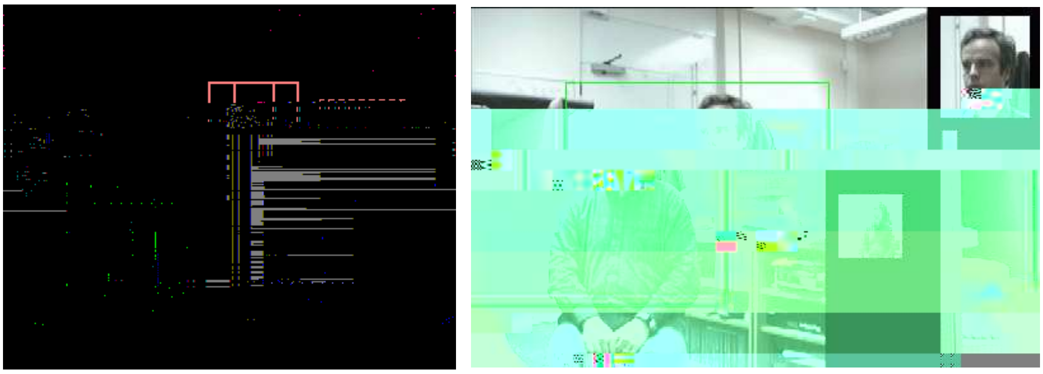
However, head pose information may be insufficient to determine




120
100
80
60
40
0
2
4
6
8
10
12

- Pre-defined gestures and head pose of several individuals in the scene can be simultaneously recognised for interpretation of the scene.
- A scene vector-to-camera control transformation can be performed via a TDRBF network, using example-based learning.

We have been able to show how multi-person activity scenarios can be learned from training examples and interpolated to obtain the same interpretation for



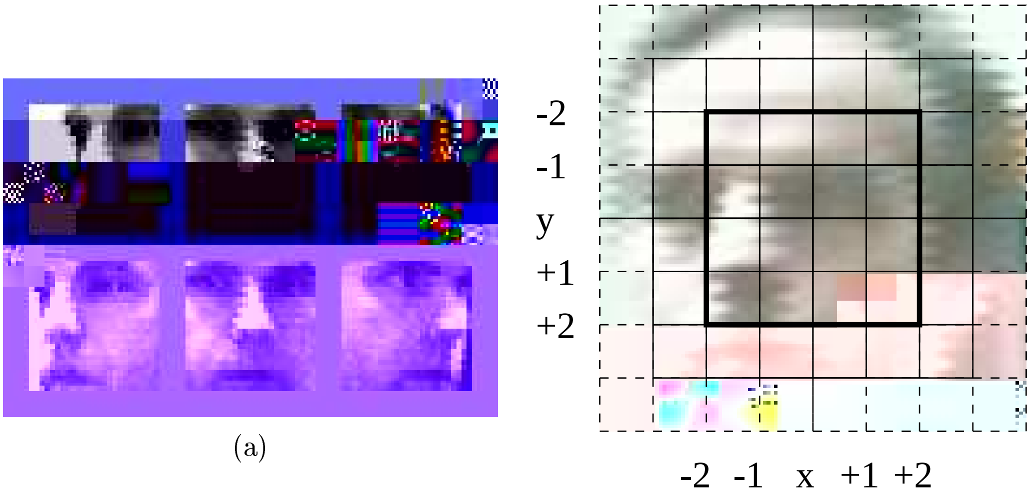
(a) (b)

Fig. 5. Use of colour/motion information to position an attentional frame around a person: (a) a box is centred around each colour/motion 'blob', the inner vertical lines representing the standard deviation of the pixels along the x-axis, giving a width measure, (b) having identified which box contains the head (the uppermost one in (a)), an attentional frame box is drawn around the person relative to the head position, and sized according to head width. The top right image shows the image area inside the head box, bottom right the resampled area of the image inside the attentional frame.

to give a binary map of moving skin pixels within the image, and we used local histogram maxima to identify potential 'blob' regions. A box which was large enough to contain the head at all distances in our target range was then fitted over the centroid of each of these regions. Fig. 5(a) shows how each box is centred on the centroid of each maximum, with the inner lines showing the standard deviation of the pixels along the x-axis from that centroid. It can also be seen that the hands are ignored in this example, as they are too low down to be included in a face-size 'blob'.

A robust approach to head tracking using colour/motion blobs is what we call *temporal matching*: the tracker only considers blobs from the current frame which have been matc28bc57centroid.usingbtface-size

fa57-a0Td[(b)-mtbnjEthinhead



(a) (b)

Fig. 6. (a) Two methods for segmenting 25×25 pose-varying face data: (top row) nose-centred, (bottom row) face-centred, the former being used for experiments here, (b) the grid system for detecting potential faces within a potential 'head blob' region of the image: each area tested is represented by a 4×4 box, the thick line shows the central position $(x, y = 0)$, normal line and dashed lines

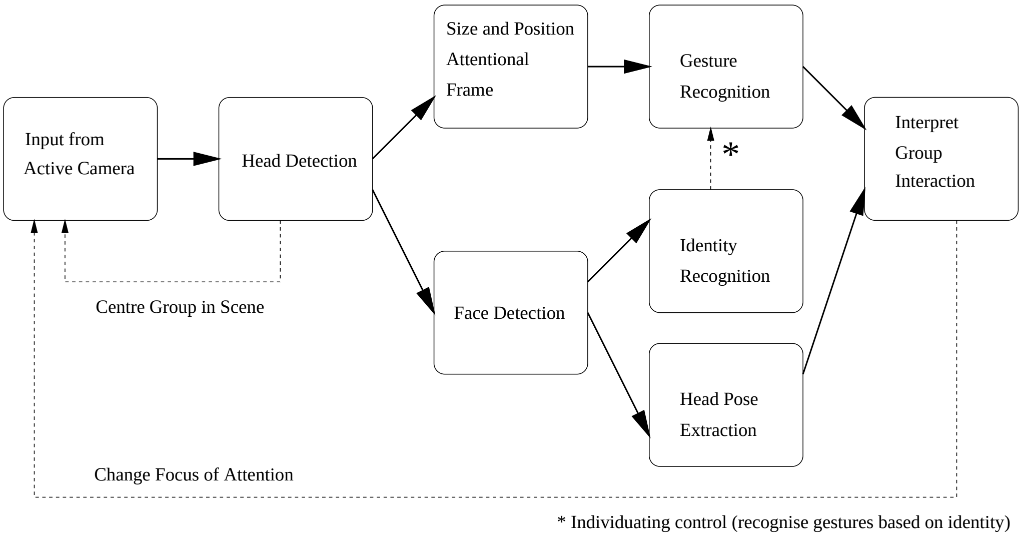


Fig. 7. A block diagram outlining the integrated system (from [22]).

qualitative level of head-pose was found to be very useful for group interaction analysis

of high-level models such as *Bayesian Belief Networks* (BBNs) might provide a combination of hand-coded *a priori* information with machine learning to ease training set requirements. This is because the BBNs model the decomposition of the problem and it is the model parameters (conditional probabilities) that are learnt so that higher level inferences can be made from low level visual evidence (see, for example, [7]).

6 Conclusions and Further Research

It is clear that there are many potential advantages of Visually Mediated Interaction with computers over traditional keyboard/mouse interfaces. For example, removing system-dependant IT training and allowing the user a more intuitive form of system direction. However, we have also seen that there are still many challenges for integrating multi-user interaction analysis and control due to the ambiguities and combinatorial explosion of possible behavioural interactions. We have demonstrated how our connectionist techniques can support real-time interaction by detecting faces and capturing 'attentional frames' to focus processing. To go further we will have to build our VMI systems around the task demands which include both the limitations of our techniques and potentially conflicting intentions from users. Connectionist techniques are generally well suited to this kind of situation as they can learn adaptive mappings and have inherent constraint satisfaction.

Further research is taking two main directions: 1) the development of gesture-based control of animated software agents in the EU Puppet project; and 2) the development of context-based control in more complex scenarios in the new EU Actipret project. The first (e.g. the GestureBall application) extends the use of symbolic (action selection) and mimetic (dynamic control) functions in gesture-based interfaces where pointing can indicate the current avatar and movement patterns can control animation parameters. The second involves recognition of complex behaviours and activities that consist of a sequence of events that evolve over time [16,17]. As yet there has been little work that combines automated learning of behaviours in different contexts. In other words, it is usually only simple, generic models of behaviour that have been learnt rather than learning when and how to apply more complex models in a context sensitive manner.

Acknowledgements

The authors gratefully acknowledge the invaluable discussion, help and facilities provided by Shaogang Gong, Jamie Sherrah and Stephen McKenna under the EPSRC-funded ISCANIT project during the development and construction of the gesture database and in collaborative work with the group interaction experiments, and also by Mike Scaife and Yvonne Rogers at the Interact Lab at the University of Sussex, for the *GestureBall* application.

References

[1] A. Blake and A. Yuille. *Active Vision*. MIT Press, 1992.

[2] A. Bobick and A. Wilson. A state-based technique for the summarization and recognition of gesture. In *Proceedings of International ionofelopmerIonthe*

[13] 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 *Proceedings of IEEE International Conference on Automatic Face & Gesture Recognition*, pages 118–123, Nara, Japan, 1998. IEEE Computer Society Press.

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

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

[16] R. J. Howarth and H. Buxton. Attentional control for visual surveillance. In S. Maybank and T. Tan, editors, *ICCV Workshop on Visual Surveillance.* IEEE Computer Society Press, 1997.

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

[18] A. J. Howell. *Automatic face recognition using radial basis function net orks.* PhD thesis, University of Sussex, 1997.

[19] A. J. Howell. Face recognition using RBF networks. In R. J. Howlett and L. C. Jain, editors, *Radial Basis Function Net orks 2: Ne Advances in Design*, pages 103–142. Physica-Verlag, 2001.

[20] A. J. Howell and H. Buxton. Learning gestures for visually mediated interaction. In P. H. Lewis and M. S. Nixon, editors, *Proceedings of British Machine Vision Conference*, pages 508–517, Southampton, UK, 1998. BMVA Press.

[21] A. J. Howell and H. Buxton. Learning identity with radial basis function networks. *Neurocomputing*, 20:15–34, 1998.

[22] A. J. Howell and H. Buxton. RBF network methods for face detection and attentional frames. *Neural Processing Letters*, 15, 2002 (In Press).

[23] M. Isaard and A. Blake. A mixed-state condensation tracker with automatic model-switching. In *Proceedings of International Conference on Computer Vision*, pages 107–112, Bombay, India, 1998. IEEE Computer Society Press.

[24] A. Jebara and A. Pentland. Action reaction learning: Automatic visual analysis and synthesis of interactive behaviour. In *Proceedings of International Conference on Vision Systems (ICVS'99)*, Las Palmas de Gran Canaria, Spain, 1999.

[25] T. Jebara and A. Pentland. On reversing Jensen's Inequality. In *Advances in Neural Information Processing Systems*, volume 13, Denver,Colorado, 2000.

[26] G. Johansson. Visual perception of biological motion and a model for its analysis. *Perception and Psychophysics*, 14:201–211, 1973.

[27] Ks [illegible]

[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, Hillsdale, NJ, 1989.

[29] L. T. Kozlowski and J. E. Cutting. Recognising the sex of a walker from a dynamic poit-light display. *Perception and Psychophysics*, 12:575–580, 1977.

[30] P. Maes, T. Darrell, B. Blumberg, and A. Pentland. The ALIVE system: Wireless, full-body interaction with autonomous agents. *ACM Multimedia Systems*, 1996.

[31] S. J. McKenna, S. Gong, and Y. Raja. Face recognition in dynamic scenes. In A. F. Clark, editor, *Proceedings of British Machine Vision Conference*, pages 140–151, Colchester, UK, 1997. BMVA Press.

[32] 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, Pittsburgh, PA, 1988. Morgan Kaufmann.

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

[34] Y. Moses, Y. Adini, and S. Ullman. Face recognition: the problem of compensating for illumination changes. In J. O. Eklundh, editor, *Proceedings of European Conference on Computer Vision, Lecture Notes in Computer Science*, volume 800, pages 286–296, Stockholm, Sweden, 1994. Springer-Verlag.

[35] M. C. Mozer. Neural net architectures for temporal sequence processing. In A. S. Weigend and N. A. Gershenfeld, editors, *Time Series Prediction: Predicting the Future and Understanding the Past*, pages 243–264. Addison-Wesley, Redwood City, CA, 1994.

[36] N. Oliver, B. Rosario, and A. Pentland. Graphical models for recognising human interactions. In *Advances in Neural Information Processing Systems*, Denver, Colorado, 1998.

[37] N. Oliver, B. Rosario, and A. Pentland. A Bayesian computer vision system for modeling human interactions. In *International Conference on Vision Systems*, Gran Canaria, Spain, 19989.

[38] A. Pentland. Smart rooms. *Scientific American*, 274(4):68–76, 1996.

[39] C. Pinhanez and A. F. Bobick. Human action detection using PNF propagation of temporal constraints. In *Proceedings of IEEE Conference on Computer Vision & Pattern Recognition*, Santa-Barbara, CA, 1998.

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

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

[42] 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, San Mateo, CA, 1989. Morgan Kaufmann.

[43] M. Rosenblum and L. S. Davis. An improved radial basis function network for autonomous road-following. *IEEE Transactions on Neural Net orks*, 7:1111–1120, 1996.

[44] J. Sherrah, S. Gong, A. J. Howell, and H. Buxton. Interpretation of group behaviour in visually mediated interaction. In *Proceedings of 15th International Conference on Pattern Recognition*, pages 266–269, Barcelona, Spain, 2000.

[45] R. H. Thibadeau. Artificial perception of actions. *Cognitive Science*, 10:117–149, 1986.

[46] M. Turk. Visual interaction with lifelike characters. In *Proceedings of International Conference on Automatic Face & Gesture Recognition*, pages 368–373, Killington, VT, 1996. IEEE Computer Society Press.

[47] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. Lang. Phoneme recognition using time-delay neural networks. *IEEE Transactions on Acoustics, Speech, & Signal Processing*, 37:328–339, 1989.

[48] A. D. Wilson and A. F. Bobick. Recognition and interpretation of parametric gesture. In *Proceedings of International Conference on Computer Vision*, pages 329–336, Bombay, India, 1998. IEEE Computer Society Press.

[49] C. R. Wren, A. Azarbayejani, T. Darrell, and A. P. Pentland. Pfinder: Real-time tracking of the human body. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 19:780–785, 1997.

[50] C. R. Wren and A. P. Pentland. Dynamic models of human motion. In *Proceedings of IEEE International Conference on Automatic Face & Gesture Recognition*, pages 22–27, Nara, Japan, 1998. IEEE Computer Society Press.