Towards Visually Mediated Interaction using Appearance-Based Models

A. Jonathan Howell and Hilary Buxton

CSRP 490

June 1998

ISSN 1350–3162

UNIVERSITY OF

SUSSEX

AT BRIGHTON

Cognitive Science
Research Papers

Towards Visually Mediated Interaction using Appearance-Based Models

A. Jonathan Howell and Hilary Buxton

School of Cognitive and Computing Sciences,
University of Sussex, Falmer, Brighton BN1 9QH, UK
{jonh,hilaryb}@cogs.susx.ac.uk

Abstract. This paper reports initial research on supporting Visually Mediated Interaction (VMI) by developing generic expression models and person-specific and generic gesture models for the control of active cameras. We investigate the recognition of both head pose and expression using simple generalisation of trained generic models using Radial Basis Function (RBF) networks. Then we go on to describe a time-delay variant (TDRBF) of the network and evaluate its performance on recognising simple pointing and waving hand gestures in image sequences. Experimental results are presented that show that high levels of performance in gesture recognition can be obtained using these techniques, both for particular individuals and across a set of individuals. Characteristic visual evidence can be automatically selected and used even to recognise individuals from their gestures, depending on the task demands.

1 Introduction

This paper reports initial research on supporting Visually Mediated Interaction (VMI) by developing generic expression and gesture models. Ultimately, these models will be used to recognise intentions for active camera control, but here we describe adaptive learning techniques to develop the models. In general, robust tracking of non-rigid objects such as human bodies is difficult due to rapid motion, occlusion and ambiguities in segmentation and model matching. Ongoing research at the MIT Media Lab has shown progress in the modelling and interpretation of human body activity [25, 31, 32]. Computationally simple view-based approaches to action recognition have also been proposed [4] and similar attempts have been made at Microsoft Research [29, 6]. However, these systems do not attempt intentional tracking and modelling to control active cameras for Visually Mediated Interaction (VMI). Previous work on vision-based camera control has been based on off-line execution of pre-written scripts of a set of defined camera actions [26]. This system used a fixed wide angle camera with virtual windows for the control of field-of-view. Here we propose to model and exploit a set of 'interaction-relevant' gestures for reactive on-line visual control. These will be interpreted as user intentions for live control of an active camera with adaptive view direction and attentional focus. In particular, pointing (for direction) and waving (for attention) are important for deliberative control and the reactive camera movements could provide the necessary visual context for applications such as group video-conferencing as well as automated studio direction.

There is growing interest in recognising human gestures from real-time video as a nonverbal modality for human-computer interaction. The main approaches involve computing low-level features from motion to form temporal trajectories that can be tracked by Hidden Markov Models or Dynamic Time Warping. However, here we explore the potential of using simple image-based differences from video sequences in conjunction with a powerful connectionist learning paradigm to account for variability in the appearance of a set of predefined gestures. The adaptive learning component is based on RBF networks, which have been identified as valuable models by a wide range of researchers [21, 1, 3]. Their main advantages are computational simplicity and robust generalisation supported by a well-developed mathematical theory. Here studies using video sequences for recognition of pointing and waving gestures involve time-delay RBFs [15] to provide fast training and on-line performance for the interactive responses required by applications with active camera control.

The main purposes of this paper are 1) to show that fast training and on-line recognition of gestures and expressions is computationally feasible, and 2) to present experimental results that show that high levels of performance for this type of gesture recognition can be obtained using these techniques both for particular individuals and across a set of individuals. Characteristic visual evidence can be automatically selected and used even to recognise individuals from their gestures, depending on the task demands.

2 The RBF and Time-Delay RBF Models

The RBF network is a two-layer, hybrid learning network [21, 22], 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.

Dynamic neural networks can be constructed by adding recurrent connections to standard multi-layer perceptrons which then form a contextual memory for prediction over time [16, 8, 24]. These partially recurrent neural networks can be trained using back-propagation but there may be problems with stability and very long training sequences when using dynamic representations. An alternative is the Time-Delay Neural Network (TDNN) model (for an introduction, see [10]), which incorporates the concept of time-delays in order to process temporal context, combining data from a fixed time 'window' into a single vector as input, see Fig. 1. The TDNN has been successfully applied to speech and handwriting recognition tasks [30]. Its structured design allows it to specialise on spatio-temporal tasks, but, as in weight-sharing network, the reduction of trainable parameters over fully connected models can increase generalisation [17]. This simple Time-Delay mechanism can be added to an RBF network, termed a TDRBF network [2], to allow fast, robust solutions to difficult real-life problems. In its original form, the TDRBF network used a constructive RBF training stage, combining the idea of a sliding input window from the standard TDNN network with a training procedure for adding and adjusting RBF units when required. We have applied a simpler technique, successful in previous work with RBF networks [13], which uses an RBF unit for each training example, and a simple pseudo-inverse process to calculate weights.

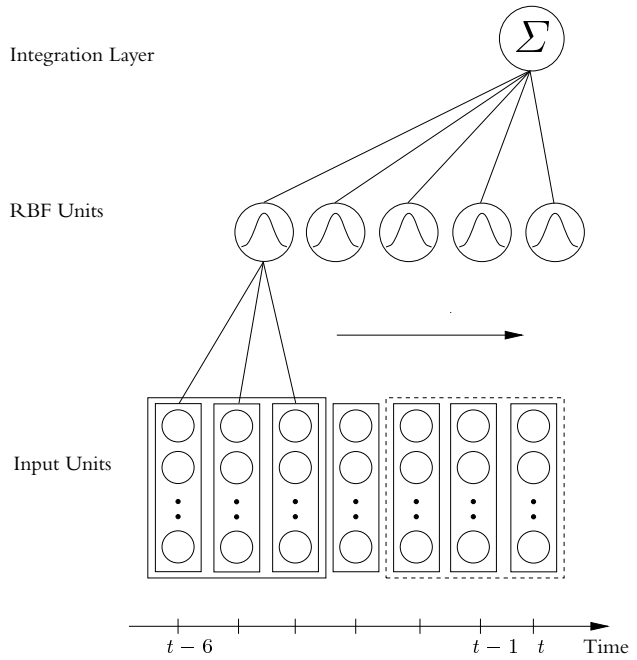


Fig. 1. Structure of a single class for a TDRBF network with time window of 3 and a integration window of 5 (after [2]).

3 Expression and Pose Experiments

The data used here was from the standard Weizmann FaceBase [23] to test identity generalisation for standard (non-TD) RBF networks in the presence of expression as well as pose variation. Although the illumination conditions can be varied, we kept it constant for these tests. For the 28 people in the database, we used 15 images for each (420 total): 5 head pose positions, *VP*, (-34°, -17°, 0°, 17°, 34°) for each of 3 expressions, *EX*, (neutral, smile, yawn).

In each table, the 'Initial % Correct' column shows the raw generalisation rate for the RBF network. This value can be adapted via a 'low-confidence' measure which has previously been shown to be a powerful method for improving generalisation [14] through the removal of ambiguous results. The '% Discarded' column indicates how many classifications were discarded in this way, and '% Correct After Discard' shows the final generalisation rate using only the resulting high-confidence output.

3.1 Expression Invariance

Our first experiment looks at identity recognition using pose- and expression-varying images. This establishes the ability of the RBF to distinguish the data in terms of identity classes before we move onto 'generic' training of the network with non-identity classes.

The 'Train/Test (per Class)' in the tables shows how many training and test images were used for each class. Following our earlier tests on pose invariance in face recognition [13], we select training data by 'paving' the pose variation found in the data for each person with equally-spaced examples (-34°, 0°, 34°), using the others for testing (-17°, 17°). This approach has been shown to provide good levels of pose invariance for trained RBF networks, and will allow us to focus on the expression variation in the data. Within this framework, then, we can then check for 'inherent' expression invariance [13] by using training examples with neutral expression only (*3 train/12 test*), and 'learnt' expression invariance by using training examples with all expressions (*9 train/6 test*).

Table 1. Test generalisation performance for RBF networks trained to classify by identity, using pose- and expression-varying images from the Weizmann FaceBase.

Train/Test (per Class)	Initial % Correct	% Discarded	% Correct After Discard
1/14	60	63	94
3/12	90	28	98
9/6	93	28	100

It can be seen in Table 1 that the RBF network can provide good identity generalisation over expression variation even without explicit examples of that variation. We also did a '1/14' test, where only one training example was used for each class. This is a harder task, as the network has to be invariant to both pose and expression, and can be compared with results in [7], where an error rate of 31% (69% correct) was reported. The major difference here is that we have not collected class-based knowledge about viewpoint and expression transformations, and therefore have only the individual training example. Such extrapolation is not a strength of RBF networks, as can been seen by the large improvement in performance when only two extra training examples per class are provided (the 3/12 test). In all cases, however, the discard measure allowed ambiguous results to be efficiently identified, as shown by high classification rates in the '% Correct After Discard' column.

3.2 Classifying Expression and Pose

In addition to identifying individuals, we also wish to develop networks for other specific tasks. With the Weizmann data, we can train RBF networks to classify images in terms of expression and pose classes, without regard for identity.

For these tests, we took all 15 images for a number of individuals to use as training examples for either the 3 expression (EX) classes or the 5 head pose (VP) classes. The 'Train/Test (Individuals)' column shows how many of these were used for training and testing. Thus, the '7/21' test used 7×15 training examples in total (split into 35 for each expression class or 21 for each pose class).

Table 2 shows that very different results were obtained for the two types of network. The head pose (VP) classes (b) appear to be very much easier to generalise to than the

Table 2. Test generalisation performance for generic RBF networks trained to classify (a) expression (EX) classes, using pose- and identity-varying images, or (b) head pose (VP) classes, using expression- and identity-varying images.

Type of Classes	Train/Test (Individuals)	Initial % Correct	% Discarded	% Correct After Discard
(a) Expression (EX)	21/7	67	41	79
(3 classes)	14/14	67	46	81
	7/21	52	53	61
(b) Head Pose (VP)	21/7	99	8	99
(5 classes)	14/14	95	19	98
	7/21	89	20	93

expression (EX) classes (a), which failed to give good results even with large numbers of training examples. This may be due to expressions requiring higher resolution data for discrimination than head pose, which can be easily identified from low resolution images.

3.3 Summary

Our initial tests with the Weizmann FaceBase show how simple RBF classifiers can be developed for particular static recognition tasks. These could be used in parallel to give a range of information about the current frame from live video input. It is clear that we need to improve on the high level of error in the undiscarded generalisation rates by using class-based knowledge, eg. [7].

4 Gesture Experiments

We now look at temporal gesture recognition in image sequences. Simple experiments have previously been made with the TDRBF network to learn behaviours based on y-axis head rotation [15], distinguishing between left-to-right and right-to-left movements and static head pose. The network was shown to maintain a high level of performance even on test data containing individuals not seen during training. However, such tasks are simplified by their constant motion, so that arbitrary short segments (2/3 frames) of the whole sequence could be used to identify the overall direction of head turning. Here, we address more complex gestures: pointing and waving with a hand. Due to the complex motion involved, all characteristic parts of the complete action will need to be contained in the time window presented to the network in order that it can be recognised.

4.1 The Gesture Database

Within the context of a video-conferencing active camera control scenario, we are concentrating on two specific behaviours which could be used to move the camera or adapt its field of view: *pointing*, which is interpreted as a request to pass camera attention, and is

implemented by zooming out and panning in the pointing direction, and *waving*, which is interpreted as a request for camera attention, and implemented by panning towards the waver and zooming in. We have two types of each behaviour, giving four gestures in all, shown in Table 3.

Table 3. Definitions for the four gestures used.

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'

We have collected four examples of each gesture from three people, 48 sequences in all, so far. Each sequence contains 59 378×288 8-bit monochrome images (having been collected at 12 frames/sec for roughly 5 seconds), for a total of 2832 images. These image sequences are the result of our collaboration in the ISCANIT project with Shaogang Gong at Queen Mary and Westfield College, London and Stephen McKenna at the University of Dundee, who are researching real-time face detection and tracking [18–20, 28]. The standard RBF and TD-RBF networks have already been shown to work well with such image sequences for face recognition tasks [14, 15].

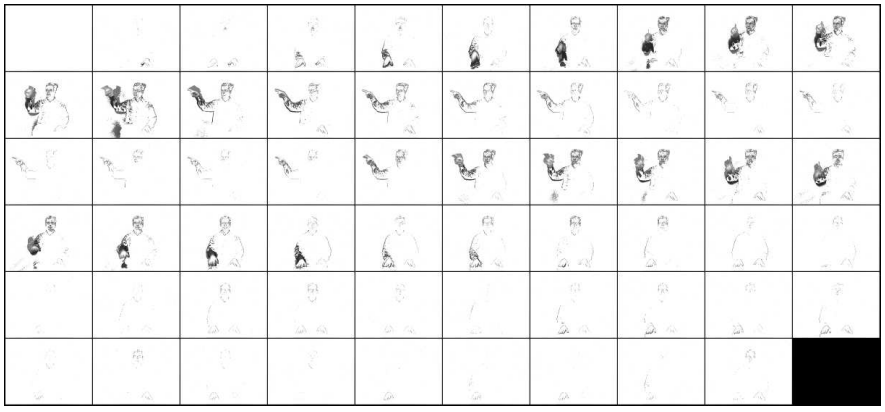
We are specifically interested in the areas of motion within each image, so each frame is differenced with the previous one: any pixel in the current frame within 5 grey-levels of the corresponding pixel from the previous frame is discarded (set to zero), see Fig. 2. A count of the number of pixels retained in each frame after this process can be used to segment the gesture in time, using a simple threshold to signal the first and last frame with significant numbers of changing pixels, see Fig. 3. Frames before and after this threshold are discarded to align the start point of the gesture. The sequences are then padded at the end with nil values to the length of the longest gesture found, to give an equal length for all sequences in the testset. An integration layer on the TDRBF network can be used to combine results from successive time windows, which will give smooth gradations between serial actions. Here we know each sequence contains only one action, and so can rely on our temporal segmentation to give the single best frame position for classification. A sparse arrangement of Gabor filters [12] is used to preprocess the differenced images: data is sampled at four non-overlapping scales and three orientations with sine and cosine components for a total of 510 coefficients per frame.

4.2 Gesture Results

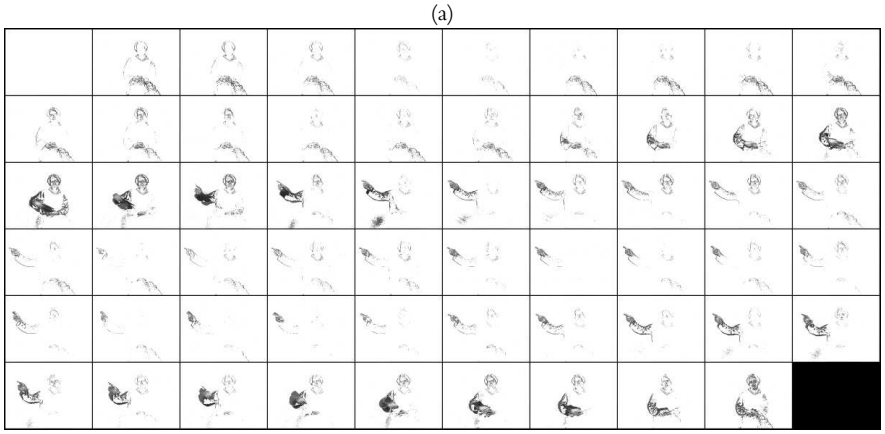
Table 4 summarises the results obtained. For all the experiments in this section, the database was split into two separate parts, one for training and the other for testing. The 'Train/Test Sequences' column shows the actual number of sequences in each part for the experiment. As before, the 'Initial % Correct' column shows the raw generalisation rate for the TDRBF network, the '% Discarded' column indicates how many classifications

(a) (b) (c)

Fig. 2. Example of differencing two consecutive frames (a) and (b), result in (c), from a 'point with the right hand to the right' (*pntrr*) gesture.



(a)



(b)

Fig. 3. Two example 5-second image sequences (after differencing each frame with the previous one) of the 'point with the right hand to the right' (*pntrr*) gesture, from two different people, demonstrating the type of variability present in a single action: (a) subject *jonh*, starting frame was automatically determined as frame 4, ending frame 32 (b) subject *step*, starting frame 18, ending frame 57.

were discarded in this way, and '% Correct After Discard' shows the final generalisation rate using only the resulting high-confidence output.

Person-specific Gesture Modelling We first looked at creating person-specific networks, trained and tested using data from one individual. For this, we used the 16 sequences of each person separately in turn. Here we are looking to distinguish the four gestures, so there are four classes to be learnt.

Table 4(a) lists the results for two training configurations, averaged over the three sets of data. These show that the task could be learnt extremely well even with only one training example of each gesture (the 4/12 test), but that providing two examples (the 8/8 test) can reduce the number of low-confidence discards, indicating a more effective separation of the gesture classes within the network.

Table 4. Average results for various TDRBF gesture networks: (a) *person-specific*, trained and tested with gesture sequences from a single person, (b) *group-based*, trained and tested with gesture sequences from three people, and (c) *generic*, trained with gesture sequences from one person, and tested on the other two.

Gesture Model	Train/Test Sequences	Initial % Correct	% Discarded	% Correct After Discard
(a) Person-specific	4/12	92	51	100
	8/8	100	16	100
(b) Group-based	12/36	97	39	100
	24/24	96	4	100
(c) Generic	4/32	69	75	100
	8/32	75	53	100

Group-based Gesture Modelling We now combine the data from the three people in the database, using all 48 sequences, looking at generalisation in the TDRBF network within a group seen during both training and testing. Again we are looking to distinguish the four gestures, so there are four classes to be learnt. In general, the results in Table 4(b) are slightly better than before: the final result is still 100%, but at lower levels of low-confidence discard, especially where two training examples of each gesture are given (the 24/24 test).

Generic Gesture Modelling Having tested the TDRBF network with single and multiple-person data, we also wanted to see how it would generalise with gestures from people it had not seen during training: could gesture be effectively characterised in the absence of identity-specific information? There are still four classes to be learnt, but the network sees examples from one person during training, and the other two during testing.

It can be seen in Table 4(c), which follows the pattern of the previous tests by using one (4/32) and two (8/32) training examples of each gesture, that the TDRBF network

was able to learn the classification task from one person and effectively generalise to data from other people, though with a higher discard rate than in the previous two experiments. Such an ability is potentially much more useful than generalisation within specific people or known groups, because the network, once trained, can be applied to much more general data. Generalisation could be improved by using more people for training [20].

Combined Gesture and Identity Modelling Our final test was to see if the individual, as well as the gesture, could be identified. The three identities and four gestures mean that there are now 12 classes to be learnt. Table 5 gives the results for these tests. When compared to the group-based results in Table 4(b), it can be seen that overall generalisation is lower, although a significant proportion of the test data is still correctly classified. This indicates that non-facial cues, such as posture and gait, could be used in this form as additional evidence for face recognition applications.

Table 5. Results for identity/gesture TDRBF networks: trained and tested with gesture sequences of all three people, looking both for identity and gesture.

Train/Test Sequences	Initial % Correct	% Discarded	% Correct After Discard
12/36	78	50	94
24/24	96	17	95

4.3 Discussion

Although Berthold [2] used the integration layer to cope with shifts in time, the scale of events was not discussed. In particular, here we have to cope with different speeds of movement and pauses within the overall gesture, as well as the starting frame of the gesture being variable. Such speed variation can be handled by a recurrent network, or via training data which explicitly demonstrated the classes at different speeds (time-warping).

To simplify the results here, an integration layer was not used during the testing stage. The 'pixels-changed' threshold, looking at overall movement within the frame, was effective within this database in identifying start and end points of gestures, but would not be robust in more general situations, especially if the scene contained more than one person. We anticipate that adding an integration layer would improve results, because the extra variation in starting point for the test sequences (through their iterative application on successive frames) would give extra contextual information for identification.

5 Observations

Several points can be seen from the results:

- Computationally efficient RBF classifiers can be developed for particular static recognition tasks: identity, expression and pose.
- Straightforward preprocessing techniques such as frame differencing and thresholding can be effective in extracting motion information for temporal segmentation of gestures.
- Several types of TDRBF network gesture model can be trained for recognition within specific time windows: *Person-specific*. trained and tested on one person *Group-based*, trained and tested within a known group of individuals, and *Generic*, trained on one person, tested on other people.
- The TDRBF network is shown to be able to distinguish between arbitrary gestures, with a high level of performance, even without the benefit of an integration layer. The temporal segmentation allowed a single time window to be applied to the network, rather than several consecutive positions.
- Some characteristics of an individual's expression of gestures may be sufficiently distinctive to identify that person.

6 Conclusion

In summary, the time-delay RBF networks showed themselves to perform well in our gesture recognition task, creating both person-specific and generic gesture models. This is a promising result for the RBF techniques considering the high degree of potential variability, present even in our highly constrained database, arising out of the different interpretation of our predefined gestures by each individual.

In our new project, we aim to develop and evaluate real-time user behaviour models based on temporal prediction of continuous pose and gesture change [9, 27]. The user would have minimal awareness of the system which will aim to estimate and predict essential body parameters such as head pose, walking, sitting, standing, talking, pointing and waving gestures as well as expression. Such a model will be essentially appearance-based in order to provide real-time behaviour interpretation and prediction. It is important to note that we are not attempting to model the full working of the human body. Rather we will aim to exploit approximate and computationally efficient RBF techniques, which support partial view-invariance, sufficient to recognise people's expressions and gestures in dynamic scenes. Such task-specific representations need to be used to avoid unnecessary computational cost in dynamic scene interpretation [5].

Most existing recurrent network models take a long time to train, but simple time-delay RBF networks provide a fast and effective method of identifying arbitrary behaviours [15]. The main problem with this alternative strategy for learning behavioural models is that it is difficult to classify the same behaviour evolving at different speeds

using a single time-window. Solutions to this problem require either a) subdividing the behaviours into fast and slower versions and/or b) merging these in a second stage of behavioural analysis. This flexibility may turn out to be an advantage in practice as the intentional force of a fast pointing action (urgent) may be different from a slower action. We therefore plan to explore the use of full generative RNNs [11] for general behavioural control that can be learnt incrementally using many examples and time-delay RBFs for individual intentional control which needs to be learnt rapidly from a few examples.

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, San Mateo, CA, 1993. Morgan Kaufmann.
2. M. R. Berthold. A time delay radial basis function network for phoneme recognition. In *Proceedings of IEEE International Conference on Neural Networks*, volume 7, pages 4470–4473, Orlando, FL, 1994. IEEE Computer Society Press.
3. C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, Oxford, UK, 1995.
4. A. F. Bobick. Computers seeing action. In R. B. Fisher and E. Trucco, editors, *Proceedings of British Machine Vision Conference*, pages 13–22, Edinburgh, 1996. BMVA Press.
5. H. Buxton and S. Gong. Visual surveillance in a dynamic and uncertain world. *Artificial Intelligence*, 78:431–459, 1995.
6. R. Cutler and M. Turk. View-based interpretation of real-time optical flow for gesture recognition. In *Proceedings of IEEE International Conference on Automatic Face & Gesture Recognition*, pages 416–421, Nara, Japan, 1998. IEEE Computer Society Press.
7. 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.
8. J. Elman. Finding structure in time. *Cognitive Science*, 14:179–211, 1990.
9. S. Gong. Visual observation as reactive learning. In *Proceedings of SPIE International Conference on Adaptive & Learning Systems*, pages 265–270, Orlando, FL, 1992.
10. J. A. Hertz, A. Krogh, and R. G. Palmer. *Introduction to the Theory of Neural Computation*. Addison-Wesley, Redwood City CA, 1991.
11. G. E. Hinton and Z. Ghahramani. Generative models for discovering sparse distributed representations. *Philosophical Transactions of Royal Society London, Series B*, 352:1177–1190, 1997.
12. A. J. Howell and H. Buxton. Receptive field functions for face recognition. In *Proceedings of 2nd International Workshop on Parallel Modelling of Neural Operators for Pattern Recognition*, pages 83–92, Faro, Portugal, 1995. University of Algarve.
13. A. J. Howell and H. Buxton. Face recognition using radial basis function neural networks. In R. B. Fisher and E. Trucco, editors, *Proceedings of British Machine Vision Conference*, pages 455–464, Edinburgh, 1996. BMVA Press.
14. A. J. Howell and H. Buxton. Towards unconstrained face recognition from image sequences. In *Proceedings of International Conference on Automatic Face & Gesture Recognition*, pages 224–229, Killington, VT, 1996. IEEE Computer Society Press.
15. A. J. Howell and H. Buxton. Recognising simple behaviours using time-delay RBF networks. *Neural Processing Letters*, 5:97–104, 1997.

16. 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.
17. Y. Le Cun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1:541–551, 1989.
18. S. J. McKenna and S. Gong. Tracking faces. In *Proceedings of International Conference on Automatic Face & Gesture Recognition*, pages 271–276, Killington, VT, 1996. IEEE Computer Society Press.
19. 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.
20. S. J. McKenna, S. Gong, and Y. Raja. Gesture recognition for visually mediated interaction using probabilistic event trajectories. In *Proceedings of British Machine Vision Conference*, Southampton, UK, 1998. BMVA Press.
21. 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.
22. J. Moody and C. Darken. Fast learning in networks of locally-tuned processing units. *Neural Computation*, 1:281–294, 1989.
23. 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.
24. 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.
25. A. Pentland. Smart rooms. *Scientific American*, 274(4):68–76, 1996.
26. C. Pinhanez and A. F. Bobick. Approximate world models: Incorporating qualitative and linguistic information into vision systems. In *AAAI'96*, pages ??–??, Portland, Oregon, 1996.
27. A. Psarrou, H. Buxton, and S. Gong. Modelling spatio-temporal trajectories and face signatures on partially recurrent neural networks. In *Proceedings of IEEE International Conference on Neural Networks*, volume 5, pages 2226–2231, Perth, Australia, 1995.
28. Y. Raja, S. J. McKenna, and S. Gong. Tracking and segmenting people in varying lighting conditions using colour. In *Proceedings of IEEE International Conference on Automatic Face & Gesture Recognition*, pages 228–233, Nara, Japan, 1998. IEEE Computer Society Press.
29. 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.
30. 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.
31. C. R. Wren, A. Azarbayejani, T. Darrell, and A. P. Pentland. Pfinder: Real-time tracking of the human body. In *Proceedings of International Conference on Automatic Face & Gesture Recognition*, pages 51–56, Killington, VT, 1996. IEEE Computer Society Press.
32. 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.