Learning Gestures for Visually Mediated Interaction

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 person-specific and generic gesture models for the control of active cameras. We describe a time-delay variant of the Radial Basis Function (TDRBF) 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 can be obtained for this type of gesture recognition 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

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 [22, 28, 29]. Computationally simple view-based approaches to action recognition have also been proposed [4] and similar attempts have been made at Microsoft Research [26, 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 [23]. 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

BMVC 1998 doi:10.5244/C.12.51

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 [19, 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 [14] to provide fast training and on-line performance for the interactive responses required by applications with active camera control.

The main purpose of this paper is 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 Time-Delay RBF Model

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 [15, 7, 21]. 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. Instead, we use a simple Time-Delay mechanism in conjunction with an RBF network, which we term a TDRBF network, to allow fast, robust solutions to difficult real-life problems. The Time-Delay Neural Network (TDNN) model (for an introduction, see Hertz et al. [9]), incorporates the concept of time-delays in order to process temporal context, and has been successfully applied to speech and handwriting recognition tasks [27]. Its structured design allows it to specialise on spatio-temporal tasks, but, as in weight-sharing network, the reduction of trainable parameters can increase generalisation [16].

The RBF network is a two-layer, hybrid learning network [19, 20], 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. A Time-Delay version of this [2] can be created by combining data from a fixed time 'window' into a single vector as input. Berthold, however took a constructive approach for the 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 used a simpler technique, successful in previous work with RBF networks [12], which uses an RBF unit for each training example, and a simple pseudo-inverse process to calculate weights.

3 Method

Simple experiments have previously been made with the TDRBF network to learn certain simple behaviours based on y-axis head rotation [14], distinguishing between left-to-right and right-to-left movements and static head pose. The network was shown to maintain a

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

Table 1: Definitions for the four gestures used.

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. In this paper, we are addressing more complex gestures: pointing and waving with a hand. Due to the complex motion involved here, 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.

3.1 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 1.

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 [17, 18, 25]. The standard RBF and TD-RBF networks have already been shown to work well with such image sequences for face recognition tasks [13, 14].

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. 1. 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. 2. 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 is used to preprocess the differenced images [11]:

Train/Test Sequences	Initial % Correct	% Discarded	% Correct After Discard
4/12	92	51	100
8/8	100	16	100

Table 2: Average results for person-specific TDRBF networks: trained and tested with gesture sequences from a single person.

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 Results

Tables 2-5 summarise the results obtained. For all the experiments, the database was split into two separate parts, one for training and the other for testing. The 'Train/Test' column shows the actual number of sequences in each part for the experiment. '% Correct' shows the raw generalisation rate for the TDRBF network. This value can be adapted via a 'low-confidence' measure which has previously been shown to be a powerful method for improving generalisation [13] 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.

4.1 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 2 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.

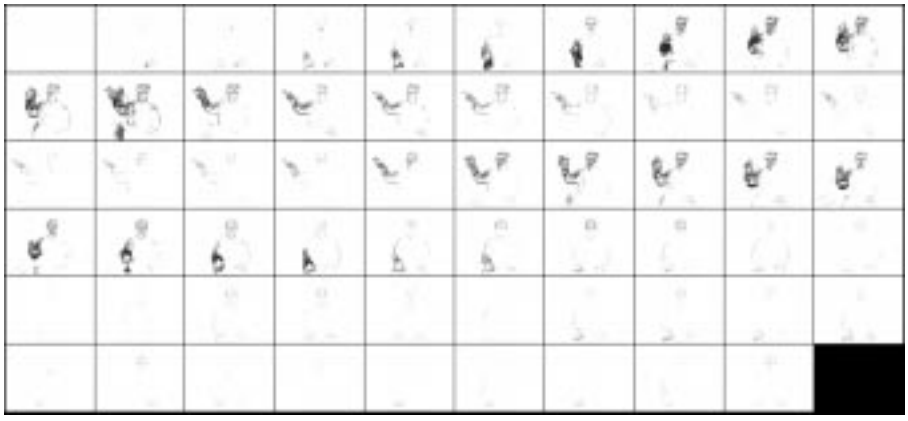
4.2 Group-based Gesture Modelling

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 3 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).

To cope with different speeds of movement, we looked at adapting the training data to explicitly demonstrate the classes at different speeds through simple time-warping. This was applied to our training sequences by cutting out or repeating frames in the time

(a) (b) (c)

Figure 1: 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)

Figure 2: 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.

Window	Train/Test Sequences	Initial % Correct	% Discarded	% Correct After Discard
Fixed	12/36	97	39	100
	24/24	96	4	100
Warped ($\pm 10\%$)	36/36	97	19	100
	72/24	96	4	100

Table 3: Results for non-person-specific TDRBF networks: trained and tested with gesture sequences from three people. Fixed window used the training sequences at their original length, the warped window used three versions (shorter, normal, longer) of the training sequences.

Train/Test Sequences	Initial % Correct	% Discarded	% Correct After Discard
4/32	69	75	100
8/32	75	53	100

Table 4: Results for generic TDRBF networks: trained with gesture sequences from one person, and tested on the other two.

window to shorten or lengthen the training sequences. Using one shorter and one longer version of each sequence meant that there were three times more training examples than for the previous experiment (with fixed time windows). The results for this are also shown in Table 3. Interestingly, varying the length by $\pm 10\%$ did not increase generalisation rates (perhaps because there was very little improvement that could be made), though it did make a useful reduction in the proportion of output that needed to be discarded through low confidence. However, it should be noted that such global methods are limited in their application, as they can only address overall gesture tempo, not within-gesture speed changes.

4.3 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, 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. 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.

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

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

4.4 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 3, 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.

4.5 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:

- Simple preprocessing techniques such as frame differencing and thresholding can be effective in extracting useful motion information and segmenting gestures in time.
- Several types of TDRBF network can be trained to distinguish gestures over specific time windows:
  - Person-specific gesture models: trained and tested on one person
  - Group-based gesture models: trained and tested within a known group of individuals

  - Generic gesture models: 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 thresholding in time of the gestures 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 [8, 24]. 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 [14]. 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 [10] 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.

Acknowledgements

The authors gratefully acknowledge the invaluable discussion, help and facilities provided by Shaogang Gong and Stephen McKenna during the development and construction of the gesture database.

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

[8] S. Gong. Visual observation as reactive learning. In *Proceedings of SPIE International Conference on Adaptive & Learning Systems*, pages 265–270, Orlando, FL, 1992.

[9] J. A. Hertz, A. Krogh, and R. G. Palmer. *Introduction to the Theory of Neural Computation*. Addison-Wesley, Redwood City CA, 1991.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] A. J. Howell and H. Buxton. Recognising simple behaviours using time-delay RBF networks. *Neural Processing Letters*, 5:97–104, 1997.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

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

[21] 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.

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

[23] C. Pinhanez and A. F. Bobick. Approximate world models: Incorporating qualitative and linguistic information into vision systems. In *Proceedings of AAAI'96*, pages 1116–1123, Portland, OR, 1996.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.