Recognising Simple Behaviours using Time-Delay RBF Networks

A. Jonathan Howell and Hilary Buxton

CSRP 456

February 1997

ISSN 1350–3162

UNIVERSITY OF

SUSSEX

AT BRIGHTON

Cognitive Science
Research Papers

Recognising Simple Behaviours using Time-Delay RBF Networks

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

Email: {jonh,hilaryb}@cogs.susx.ac.uk

February 1997

Abstract

This paper present experiments using an radial basis function variant of the time-delay neural network with image sequences of human faces. The network is shown to be able to learn simple behaviours based on y-axis head rotation and generalise on different data. The network model's suitability for future dynamic vision applications is discussed.

Keywords: RBF Networks, Time-Delay Networks, Vision, Temporal Behaviours, Face Recognition, Image Sequences, View Invariance.

1 Introduction

Recognising simple behaviours is an important capability for many computer vision applications, e.g. visual surveillance (Gong & Buxton 1995) or biomedical sequence understanding (Psarrou & Buxton 1993). The behaviour in the experiments reported in this paper is simply head rotation to the left or right. However, the work raises important issues for connectionist techniques: 1) time, 2) representation, and 3) learning with generalisation. Multi-layer perceptrons with supervised learning are very popular for applications which use static representations, but time is important in many domains e.g. vision, speech and motor control. Dynamic neural networks can be constructed by adding recurrent connections to form a contextual memory for prediction in time (Jordan 1989, Elman 1990, Mozer 1993). 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 simple Time Delay (TD) in conjunction with Radial Basis Function (RBF) networks to allow fast, robust solutions to our problem of recognising head turning behaviour.

The RBF network has been identified as valuable model (Moody & Darken 1988, Ahmad & Tresp 1993, Bishop 1995) and is seen as ideal for practical vision applications

by Girosi (1992) where handling sparse, high-dimensional data (common in images) and using approximation (rather than interpolation) is important for dealing with noisy, real-life data. In our previous work we have used an adaptive learning component based on RBF networks to tackle the unconstrained face recognition problem (Howell & Buxton 1996*a*) and to identify appropriate receptive field functions for this task (Howell & Buxton 1995*a*, Howell & Buxton 1995*c*). In learning to recognise behaviour with a TDRBF network, it is again important to use an input representation (now ordered in time) that allows generalisation over variations in lighting, scale and shift. From our recent work it seems that complex 2D Gabor filters (Daugman 1988), which approximate the receptive fields of simple cells in the primary visual cortex, provide just such a representation. The main purpose of this paper is to show how we can adapt this work on face recognition from a single image frame to the problem of behaviour recognition in extended video sequences. With our approach, images containing pre-segmented faces in a typical motion sequence can be analyzed to obtain the appropriate Gabor representation for each time frame in the motion sequence.

2 The Time-Delay RBF Model

The Time-Delay Neural Network (TDNN) model (for an introduction, see Hertz et al. (1991)), incorporates the concept of time-delays in order to process temporal context, and has been successfully applied to speech and handwriting recognition tasks (Waibel et al. 1989). 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 (Le Cun et al. 1989).

The RBF network is a two-layer, hybrid learning network (Moody & Darken 1988, Moody & Darken 1989), with a supervised layer from the hidden to the output units, and an unsupervised layer, from the input to the hidden units, where individual radial Gaussian functions for each hidden unit simulate the effect of overlapping and locally tuned receptive fields. The Time-Delay version of this, such as used by Berthold (1994), combines data from a fixed time 'window' into a single vector as input (see Figure 1). Berthold, however took a constructive approach, 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 (Howell & Buxton 1996*a*), which uses an RBF units for each example, and a simple pseudo-inverse process to calculate weights.

3 Application of TDRBF Model

Simple experiments were made with the TDRBF network using image sequences to train it to identify types of y-axis rotation. The data used is of 10 people each in 10 different poses at 10° intervals from face-on to profile, for details see Howell & Buxton (1995*a*). For the following tests, half of the database were used to train the network, and the other half used to test it. Two schemes were devised to split the data up: the *Alternate Frames Tests* (see Section 3.1) used alternate frames from each person, so that

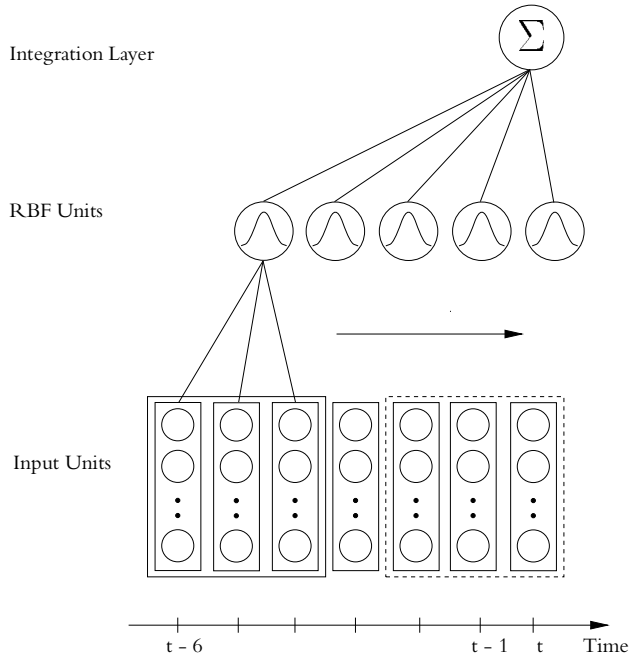

Figure 1: Structure of a single class for a TDRBF network with time window of 3 and a integration window of 5 (after Berthold (1994)).

(a) (b)

Figure 2: Example Data for Alternate Frame Tests with a Time Window of 3 Frames: (a) Training - Frames 2, 4 and 6 (b) Test - Frames 3, 5 and 7.

(a) (b)

Figure 3: Example Data for Alternate Person Tests with a Time Window of 3 Frames: (a) Training (b) Test - both using Frames 2, 3 and 4.

training and test data contained all ten people, and the *Alternate Person Tests* (see Section 3.2) used all the frames from 5 people for training, and the other 5 for testing.

Gabor wavelet analysis at a range of scales was used for preprocessing of the images. Data was sampled at four non-overlapping scales from 8×8 to 1×1 and three orientations ($0°$, $120°$, $240°$) with sine and cosine components (details in Howell & Buxton (1995*b*)). The Samples column in the tables show the total number of Gabor coefficients contained in each input vector. A discard measure was used on some of the tests to exclude low-confidence output; the proportion discarded and the subsequent generalisation rate are shown for these tests.

3.1 Alternate Frame Tests

These tests used alternate frames from all ten people for training and testing. Three types of network training were used:

Static/LR Here the training simulates left to right y-axis head rotation, and trains with a window from frames 0, 2, 4, 6 and 8 of all ten people, and tests on a window

Window	Samples	Train/Test	Initial %	% Discarded	% after Discard
5	2550	20/20	100	5	100
4	2040	40/40	95	5	100
3	1530	60/60	100	8	100
2	1020	80/80	90	8	92

Table 1: Static/LR or Static/RL Sequences from Alternate Frames (2 Classes).

Window	Samples	Train/Test	Initial %	% Discarded	% after Discard
5	2550	30/30	100	7	100
4	2040	60/60	97	8	100
3	1530	90/90	93	8	100
2	1020	120/120	83	25	96

Table 2: Static/LR/RL Sequences from Alternate Frames (3 Classes).

from frames 1, 3, 5, 7 and 9, ie using 20° intervals. Two classes are trained for: left to right movement and static. Static sequences are simulated by repeating the middle frame of the time window.

Static/RL This is similar to LR, except that the rotation is in the other direction, so that it trains with frames 8, 6, 4, 2 and 0, and tests on 9, 7, 5, 3 and 1.

Static/LR/RL This is similar to LR and RL, but trains for three classes: left to right movement, right to left movement and static.

As the LR sequence vectors are mirror-images of the RL in Euclidean space, ie the distance of LR sequence 3-5-7 to 2-4-6 is the same as the RL sequence 7-5-3 to 6-4-2, the results for the Static/RL tests are identical to those for the Static/LR tests.

3.2 Alternate Person Tests

These tests used alternate people for training and testing, each using data from five people. This is a harder test for the network, as it is tested with images of people not seen during training. Three types of network training were used:

Static/LR As before, but trains with a window from all ten frames from 0 to 9 of five people, and tests on a window from all ten frames from 0 to 9 of the other five, ie using 10° intervals.

Static/RL As before, but trains and tests with frames from 9 to 0.

Static/LR/RL As before.

Window	Samples	Train/Test	Initial %	% Discarded	% after Discard
10	4410	10/10	90	40	100
8	3528	30/30	100	30	100
6	2646	50/50	98	22	100
4	1764	70/70	91	29	98
2	882	90/90	81	40	89

Table 3: Static/LR or Static/RL Sequences from Alternate People (2 Classes).

Window	Samples	Train/Test	Initial %	% Discarded	% after Discard
10	4410	15/15	87	20	100
8	3528	45/45	89	13	100
6	2646	75/75	83	21	100
4	1764	105/105	77	37	100
2	882	135/135	63	56	93

Table 4: Static/LR/RL Sequences from Alternate People (3 Classes).

Figure 4: The test image sequence. Note the variation in head position and gaze direction.

Window	Samples	Training/Test	Integration Layer				
			1	3	5	7	9
6	3060	75/57	54	53	53	53	54
5	2550	90/58	62	62	67	64	69
4	2040	105/59	64	61	76	83	75
3	1530	120/60	63	60	73	80	78
2	1020	135/61	56	56	52	57	48

Table 5: Static/LR/RL Sequences From Alternate People (Tested on QMW Sequence).

4 Use of Real Image Sequences

To investigate the TDRBF network further, trained networks were tested on previously unseen image sequences containing a variety of head movement (see Figure 4). These image sequences are the result of collaboration with Stephen McKenna and Shaogang Gong at Queen Mary and Westfield College (QMW), University of London, who are researching real-time face detection and tracking. The standard RBF network has already been shown to work well with this data (Howell & Buxton 1996*b*).

The issue of the *time base* of actions, ie how fast or slow actions occur, was seen to be important here. Although Berthold (1994) 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 head rotation. This type of variation can be handled by a recurrent network, or training data which explicitly demonstrated the classes at different speeds. Taking this into account, the original image sequence was subsampled to match the rotation speed of the original data, which was 10° per time step. An integration layer was introduced for this test, as the network was being tested on a real image sequence. The optimum size for this layer seems to be around 7 time steps, reflecting the slow speed of head rotation present in the data.

5 Observations

Several points can seen from the results:

- The TDRBF network is shown to be able to learn certain simple behaviours based on y-axis head rotation.
- The TDRBF network maintained a high level of performance even on data containing individuals not seen during training (the alternate person test).
- An integration layer in a TDRBF network can allow the extraction of behaviour information even with quite markedly different data to that with which the network was trained.

6 Conclusion

The main points here are 1) the simple, deterministic 'training' of the TDRBF networks means that they are highly suited to on-line learning, 2) the shift invariance and ability to recognise features in time means they are capable of recognising simple behaviours, and 3) high levels of performance on the generalisation to new datasets that behave in similar ways means they are very useful for such practical dynamic vision tasks. The limitations of this technique are 1) the problem of the time-base which was not fully overcome even with the addition of an integration layer, and 2) the problem of defining the simple behaviours. The TDRBF networks are capable of distinguishing a 'quick turn' from a 'slow turn' as well as distinguishing whether the turn was to the right or the left, but it seems that more qualitative definitions of behaviour would best be tackled using more general recurrent networks. This issue is discussed further by Mozer (1993) and by Psarrou & Buxton (1994). In addition, Cleeremans (1989) shows that partially recurrent networks together with a qualitative input representation can be sucessfully used even for the demanding task of predicting state to state transitions in finite state automata. It is clear, however, that the TDRBF networks are able to perform extremely well where there is a straightforward quantitative relationship between the data and the simple behaviour pattern to be learnt.

References

Ahmad, S. & Tresp, V. (1993), Some solutions to the missing feature problem in vision, *in* S. J. Hanson, J. D. Cowan & C. L. Giles, eds, 'Advances in Neural Information Processing Systems', Vol. 5, Morgan Kaufmann, pp. 393–400.

Berthold, M. R. (1994), A time delay radial basis function network for phoneme recognition, *in* 'Proceedings of International Conference on Neural Networks', Vol. 7, Orlando, pp. 4470–4473.

Bishop, C. M. (1995), *Neural Networks for Pattern Recognition*, Oxford University Press.

Cleeremans, A. (1989), 'Finite state automata and simple recurrent networks', *Neural Computation* **1**, 372–381.

Daugman, J. G. (1988), 'Complete discrete 2-D gabor transforms by neural networks for image analysis and compression', *IEEE Transactions on Acoustics, Speech, & Signal Processing* **36**, 1169–1179.

Elman, J. (1990), 'Finding structure in time', *Cognitive Science* **14**, 179–211.

Girosi, F. (1992), 'Some extensions of radial basis functions and their applications in artifical intelligence', *Computers & Mathematics with Applications* **24**(12), 61–80.

Gong, S. & Buxton, H. (1995), Advanced visual surveillance using bayesian nets, *in* 'IEEE Workshop on Context-Based Vision', Cambridge, MA.

Hertz, J. A., Krogh, A. & Palmer, R. G. (1991), *Introduction to the Theory of Neural Computation*, Addison-Wesley.

Howell, A. J. & Buxton, H. (1995*a*), 'Invariance in radial basis function neural networks in human face classification', *Neural Processing Letters* **2**(3), 26–30.

Howell, A. J. & Buxton, H. (1995*b*), Receptive field functions for face recognition, *in* 'Proceedings of 2nd International Workshop on Parallel Modelling of Neural Operators for Pattern Recognition', University of Algarve, Faro, Portugal, pp. 83–92.

Howell, A. J. & Buxton, H. (1995*c*), A scaleable approach to face identification, *in* 'Proceedings of International Conference on Artificial Neural Networks', Vol. 2, EC2 & Cie, Paris, France, pp. 257–262.

Howell, A. J. & Buxton, H. (1996*a*), Face recognition using radial basis function neural networks, *in* 'Proceedings of British Machine Vision Conference', BMVA, Edinburgh, pp. 455–464.

Howell, A. J. & Buxton, H. (1996*b*), Towards unconstrained face recognition from image sequences, *in* 'Proceedings of International Conference on Automatic Face & Gesture Recognition', IEEE Computer Society Press, Killington, VT, pp. 224–229.

Jordan, M. (1989), Serial order: A parallel distributed processing approach, *in* 'Advances in Connectionist Theory', Erlbaum.

Le Cun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. & Jackel, L. D. (1989), 'Backpropagation applied to handwritten zip code recognition', *Neural Computation* **1**, 541–551.

Moody, J. & Darken, C. (1988), Learning with localized receptive fields, *in* D. Touretzky, G. Hinton & T. Sejnowski, eds, 'Proceedings of 1988 Connectionist Models Summer School', Morgan Kaufmann, Pittsburg, PA, pp. 133–143.

Moody, J. & Darken, C. (1989), 'Fast learning in networks of locally-tuned processing units', *Neural Computation* **1**, 281–294.

Mozer, M. (1993), Neural net architectures for temporal sequence processing, *in* A. Weigend & N. Gershenfeld, eds, 'Time Series Prediction: Predicting the Future and Understanding the Past', Addison-Wesley.

Psarrou, A. & Buxton, H. (1993), 'Hybrid architecture for understanding motion sequences', *Neurocomputing* **5**, 221–241.

Psarrou, A. & Buxton, H. (1994), Motion analysis with recurrent neural nets, *in* 'Proceedings of International Conference on Artificial Neural Networks', Sorrento, Italy, pp. 54–57.

Waibel, A., Hanazawa, T., Hinton, G., Shikano, K. & Lang, K. (1989), 'Phoneme recognition using time-delay neural networks', *IEEE Transactions on Acoustics, Speech, & Signal Processing* **37**, 328–339.