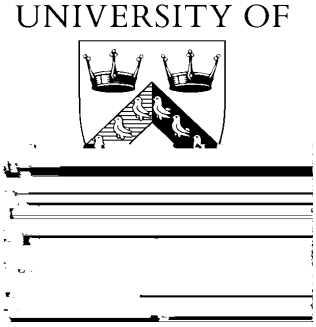
Gesture Recognition for Visually Mediated Interaction

A. Jonathan Howell and Hilary Buxton

CSRP 514

November 1999

ISSN 1350–3162



Cognitive Science

[illegible]

A. Jonathan Howell and Hilary Buxton

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

A [illegible] – This paper reports initial research on supporting Visually Mediated Interaction (VMI) by developing person-specific and generic gesture models for the cones6m1.b.652(o)0.99(s)2(t)-1.3346m1.f21433312(o)0.9966538(r)21.p6658(a)2.6665larstcnade

is based on RBF networks, which have been identified as valuable models by a wide

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, 23]. 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 [29]. 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.

o

Simple experiments have previously been made with the TDRBF network to learn certain simple 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

– 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

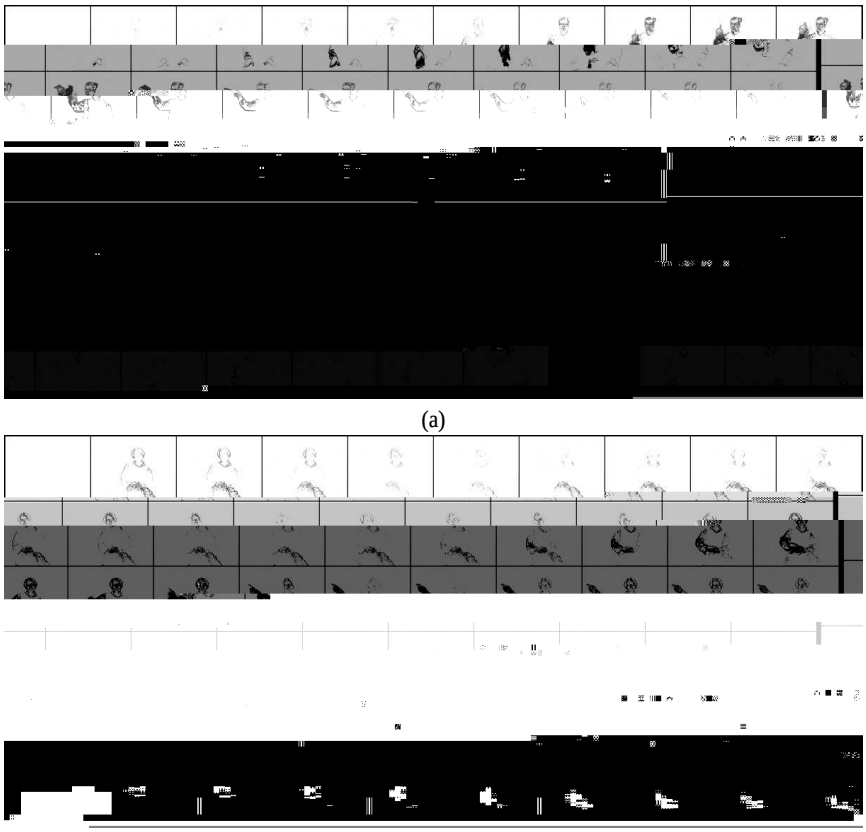
Gong at Queen Mary and Westfield College, London and Stephen McKenna at the University of Dundee, who are researching real-time face detection and tracking [18, 20, 19, 27]. The standard RBF and TD-RBF networks have already been shown to work well with such image sequences for face recognition tasks [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 is used to preprocess the differenced images [12]: data is sampled at four non-overlapping scales and three orientations with sine and cosine components for a total of 510 coefficients per frame.

Table 2 summarises the results obtained. For all the experiments in this paper, 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. ‘Initial % Correct’ shows the raw ge8535(t)4(o)-41(/)2(fi5(t)4(s)-471.992(i))ts wsher th.

(a) (b) (c)

F, –.– 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)

F, – – 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.

Table 2(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.

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: 'Fixed window' used the training sequences at their original length, 'Warped window' used three versions (shorter, normal, longer) of the training sequences, 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: Fixed	12/36	97	39	100
	24/24	96	4	100
Warped ($\pm 10\%$)	36/36	97	19	100
	72/24	96	4	100
(c) Generic	4/32	69	75	100
	8/32	75	53	100

Group-based recognition

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 2(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).

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 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 2(b). 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.

– n r, - r o ,n

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.

The test results in Table 2(c) follow the previous pattern of using one (4/32) and two (8/32) training examples. It can be seen here 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.

– Co- ,n - r n n, o ,n

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 3 gives the results for these tests. When compared to the group-based results in Table 2(b), it can be seen that overall generalisation is lower, although a significant propor

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.

Discussion

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.

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, 26]. 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] fora2(t)4(09.993(g)2.11(r)-1m9.993(g)-210.007(t535

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.