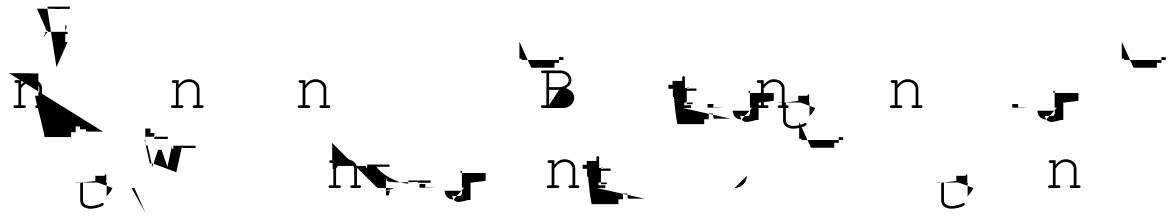

**A. Jonathan Howell and Hilary Buxton**

# Invariance in Radial Basis Function Neural Networks in Human Face Classification

A. Jonathan Howell and Hilary Buxton
School of Cognitive and Computing Sciences, University of Sussex,
Falmer, Brighton BN1 9QH, United Kingdom
{jonh,hilaryb}@cogs.susx.ac.uk

February 1995

**Abstract**

This paper is concerned with the types of invariance exhibited by Radial Basis Function (RBF) neural networks when used for human face classification, and the generalisation abilities arising from this behaviour. Experiments using face images in ranges from face-on to profile are presented to show the RBF network's invariance to 2-D shift, scale and $y$-axis rotation. Finally, the suitability of RBF techniques for future, more automated face classification purposes is discussed.

## 1 Introduction

The work reported here is based on a masters degree project and dissertation (Howell 1993) completed at the University of Sussex, and is primarily concerned with the generalisation capabilities exhibited by RBF neural networks in a static face recognition task. This work is being extended to a more fully 'environmental' process which is able to identify individuals in video sequences and interpret their gestures. The human face poses several severe tests for any visual system: the high degree of similarity between different faces, the extent to which expressions and hair can alter the face, and the large number of angles from which a face can be viewed in common situations. A face recognition system must be robust with respect to this variability and generalise over a wide range of conditions to capture the essential similarities for a given human face.

The RBF network has been identified as valuable model by a wide range of researchers (Moody & Darken 1988, Moody & Darken 1989, Poggio & Girosi 1990, Girosi 1992, Musavi et al. 1992, Ahmad & Tresp 1993). Its main char-

by a well-developed mathematical theory (resulting in statistical robustness). RBFs are seen as ideal for practical vision applications by Girosi (1992) as they are good at handling sparse, high-dimensional data (common in images), and because they use approximation which is better than interpolation for handling noisy, real-life data. RBF networks are claimed to be more accurate than those based on Back-Propagation (BP), and they provide a guaranteed, globally optimal solution via simple, linear optimisation. RBF techniques should be well suited to the face recognition task and may find second-order (relative distance) differences that can generalise well rather than first-order (absolute distance) information.

Many cognitive studies of the way human faces are perceived (Bruce & Young 1986, Bruce 1988, Ellis & Young 1989, Hay & Young 1982, Hay et al. 1991) have contributed to our understanding of the problems for automating this kind of visual processing. For example, the disproportionate effect on face recognition of inversion has been taken as support for special mechanisms in face processing (see (Hay & Young 1982) for a critical review). At a general level, there is also support for treating face classification as a task separate from, say, expression interpretation (Ellis & Young 1989). This study describes evidence for separate mechanisms being present in human vision for facial recognition and facial expression recognition. This is shown most clearly in prosopagnostic people, who cannot distinguish individual faces, but can usually still 'read' emotional states from expressions. At a more detailed level, there is support for having face 'units' for recognising familiar faces (Bruce & Young 1986, Bruce 1988). This idea is partly captured by the RBF techniques described next where the first layer of the network maps the inputs with a hidden unit devoted to each view of the face to be classified. The second layer is then trained to combine the views so that a single output unit corresponds to the individual person.

## 2 he RBF Network Model

The RBF network is a two-layer, hybrid learning network (Moody & Darken 1988), similar to the BP model in terms of structure, activation and gradient descent methods in its supervised layer from the hidden to the output nodes. However, the unsupervised layer, from the input to the hidden, differs in that individual radial Gaussian functions for each hidden unit

Each hidden unit has an associated $\sigma$ (sigma) 'width' value which defines the nature and scope of the unit's receptive field response[1]. This means that, unlike the BP network, the RBF has an activation that is related to the relative proximity of the test data to the training data. This allows a direct measure of confidence in the output of the network for a particular pattern. If a pattern is extremely different to those trained, very low (or no) output will occur.

The output $o$ for hidden unit $h$ (for a pattern $l$) can be expressed as:

$$o_h(l) = \exp[-\frac{r_h(l)^2}{\sigma_h^2}], \tag{1}$$

where

$$r_h(l) = |\mathbf{i}(l) - \mathbf{c}_h|. \tag{2}$$

The hidden layer output is also unit-normalised as suggested by (Hertz et al. 1991).

For output unit $i$, the output is:

$$o_i(l) = \sum_h w_{ih} o_h(l). \tag{3}$$

Weight adjustment is made with the Widrow-Hoff delta learning rule[2] to minimize the error measure (cost function) $\mathcal{E}$ of the network:

$$\mathcal{E} = \sum_l \mathcal{E}(l) = \sum_l \sum_i [t_i(l) - o_i(l)]^2, \tag{4}$$

where $t_i(l)$ is the target output for unit $i$ with pattern $l$.

Convergence of the network whilst training is defined as the point when the error measure for the network goes below a pre-determined 'error limit' value.

The error $\delta$ for output unit $i$ is:

$$\delta_i(l) = t_i(l) - o_i(l). \tag{5}$$

This is combined with two fixed parameters which control the speed of change, $\eta$, the learning rate, and $\alpha$, a momentum term, to give the change in value for weight $w_{ih}$ between the output and hidden layers:

$$\Delta w_{ih}(l) = \eta \delta_i(l) \sigma_h(l) + \alpha \Delta w_{ih}(l-1) \tag{6}$$

The RBF network's success in approximating non-linear multidimensional functions is dependent on sufficient hidden units being used and the suitability of the centres' distribution over the input vector space (Chen et al. 1991). In this implementation, each hidden unit centre has been set to one of the training patterns, and the weights $w_{ih}$ are initialised to the target output values, ie $w_{ih} = t_i(l)$, as recommended in Hertz et al. (1991).

[1] it is equivalent to the standard deviation of the width of the Gaussian response, so larger values allow more points to be included

[2] also known as LMS (least mean square) rule

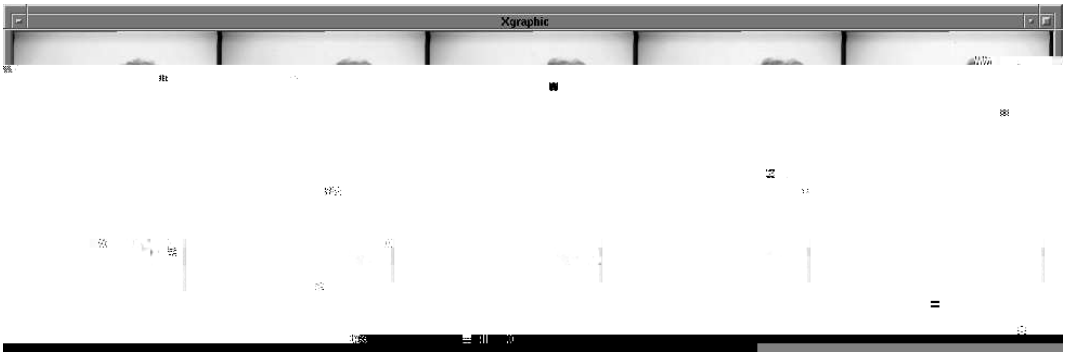


Figure 1: Entire 10-image range for one person as produced by frame grabber

# 3 Method

To simplify the problem, lighting and location for the training and test face images in these initial studies was kept as constant as possible. For each individual to be classified, ten images of the head and shoulders in ten different positions in 10° steps from face-on to profile of the left side (see Fig. 1), 90° in all, were used.

Two data sets were used: Type I with two faces, ie 20 images, for quick processing to give a general view of the networks' properties, and Type II with ten faces to give a more realistic test of the network. The resolution of the images used in the testing is represented as '$n \times n$', ie '10×10' for 10 by 10 pixel data. The ratio of training and test images used from the data set is represented as 'train/test', for instance, '2/18', where 20 images were in the data set and 2 were used for training and 18 for test.

## 3.1 Pre-processing of the Test Data

The images were gathered using a video camera and frame grabber, giving 8-bit grey-scale 384×287 images. To produce data suitable for the network, a 100×

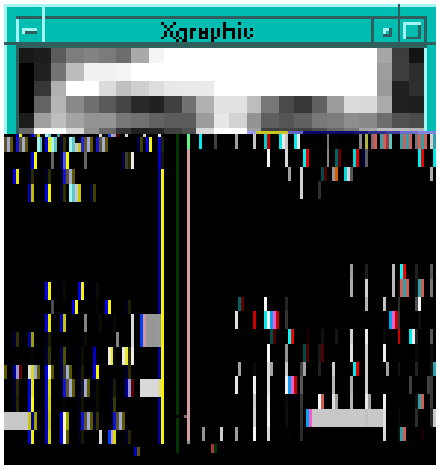


Figure 2: Example 25×25 subsampled face data

### 3.2 Image Resolution Data Sets

A range of resolutions were used for testing (the figures in brackets indicate the resolution before convolution): 10×10 (12×12), 21×21 (25×25), 44×44 (50×50), and 90×90 (100×100). If the 10×10 could give as accurate results as the 90×90, one could take advantage of a considerable reduction in the amount of data to be processed: from 8100 elements per image for 90×90 to 100 elements for the 10×10. This would be especially useful in the training stage, as the number of input units will be directly related to the computational work done, from a fraction of a second for 10×

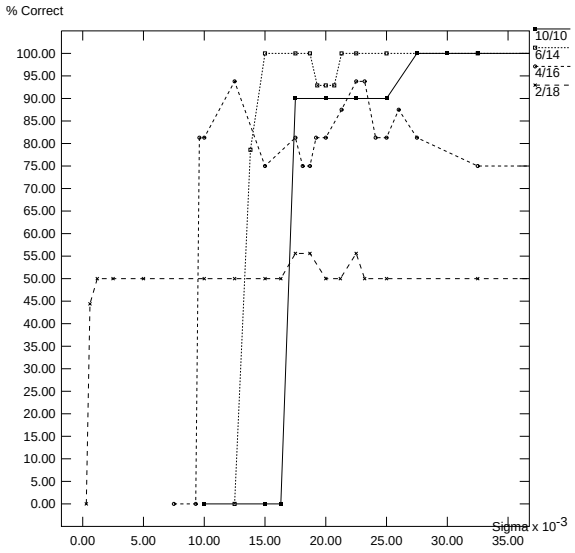


Figure 4: Comparison of generalisation to $\sigma$ with 90×90 face samples

$$\sigma_\alpha = \frac{1}{\sqrt{2}} \langle (\mathbf{c}^\alpha - \mathbf{c}^\beta)^2 \rangle^{1/2} \tag{7}$$

## 4.2 Type I Data with Varying Error Limit

The next experiment was to test the network's performance after each training epoch to show how effective a criterion the error limit was. The error limit, ie how well it could classify the training patterns, was compared with its success at classifying the test patterns.

Fig. 5, showing the 10/10 training, should be read from right to left. The network was tested to compare its success rate of classification against its current error measure value. There is a clear correspondence between recall (for the training images) and generalisation (for the test images).

2/18 training was also tested, but was much more erratic. The best performance in generalisation (88%) was early on in training at a relatively high

**Test 6:** 21×21 face samples, size variance

| Training/Test Patterns | Min/Max Epochs | Min/Max % Correct |
|---|---|---|
| | | |

seen as a smoothing factor on the energy landscape, as constructed from the energy function in Eq. 4, which helps speed the BP network gradient descent to a suitable global minimum. It is significant that the RBF network did not require this smoothing, indicating a more robust model.

The BP network with 21×21 data was capable of 100% success rates even with 2/18 training, although this dropped to 78% without added noise.

For the Type II (10 person) data, training tests with 21×21 images were attempted with the BP network using a wide range of values for the variable parameters, such as learning rate, noise level and number of hidden units, but a combination which allowed the network to converge was not found.

# 6 Conclusion/Future Work

In summary, the locally-tuned linear Radial Basis Function (RBF) networks showed themselves to be superior for the face recognition task when compared to the more complex, non-linear Back-Propagation (BP) based networks and tested on the larger 10-person data set. The RBF nets continued to show a fair level of discrimination between the different people's faces whereas the BP nets were unable to classify them at all with this data set despite producing good performance on the 2-person data set. This is a promising result for the RBF techniques considering the high degree of variability introduced by the varying views of a person's face in these data sets. The result is also backed up by the high level of performance of the RBF nets which held up with increased size and offset variance on the 2-person data set. The rather lower level of performance of the RBF nets on the 10-person data set was also little affected by the increased size and offset variance introduced in more challenging tests of their discrimination ability.

The idea of centering the nose of the profile views seems to have worked well in this study and coped with missing features from the other side of the face. This is in good accord with known results from Ahmad & Tresp (1993) who trained a variety of nets to recognise stationary hand gestures from computer-generated 2-D views (polar coordinates) of fingertips. They obtained good generalisation for 3-D orientation and showed that RBF nets were able to cope well even when much of the data was missing. Although their standard test data was handled well by a BP net, it performed badly with missing features and suffered a serious falling off in performance as more elements were lost. They showed, however, that a Gaussian RBF net (of the kind we used in our studies) could cope well, having a success rate of over 90% even with 50% of the features missing. This behaviour is very useful for coping with occlusion and other factors which lead to incomplete visual data.

The invariance observed for the RBF nets would certainly be adequate for coping with data isolated by an automated 'face-finder' routine. This is neces-

sary for the next stage of development in which faces must be found in image sequences using a combination of a wide-angle camera and separate 'retinal' camera to capture close-up views of a person's face on demand. The statistical nature of the information successfully captured by RBF nets to do the discrimination task may well also be effective for the face localisation task. It is clear from the work of Turk & Pentland (1991) and others using statistically based techniques that this is the key to good performance and the RBF techniques have the added advantage of being mathematically well-founded.

In future experiments, the performance could be improved by taking a more sophisticated measure of confidence from the output nodes in the RBF nets. For example, the simple 'winner takes all' strategy used here gave a conventional

Bruce, V. & Young, A. (1986), 'Understanding face recognition', *British Journal of Psychology* **77**, 305–327.

Chen, S., Cowan, C. F. N. & Grant, P. M. (1991), 'Orthogonal least squares learning algorithm for radial basis function networks', *IEEE Transactions on Neural Networks* **2**