Invariance in Radial Basis Function Neural Networks in Human Face Classification

A. Jonathan Howell and Hilary Buxton

CSRP 365

February 1995

ISSN 1350-3162

UNIVERSITY OF
SUSSEX
AT BRIGHTON

Cognitive Science
Research Papers

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 characteristics are first, its computational simplicity (only one layer involved in supervised training which gives fast convergence), and second, its description

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 The 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 simulate the effect of overlapping and locally tuned receptive fields. They use the vector norm distance, $r = |\mathbf{i} - \mathbf{c}|$ (which is equivalent to $r = \sum_{x=1}^{N}(i_x - c_x)^2$) between the N-dimensional input vector $\mathbf{i}$ and hidden unit centre $\mathbf{c}$ (N being the number of input units). The output value can be seen to reach a maximum when $\mathbf{i}$ nears $\mathbf{c}$. The input vectors are unit-normalised.

Each hidden unit has an associated σ (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 δ 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, η, the learning rate, and α, 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 100×100-pixel 'window' from which to take face information was located manually in each image. This was centred on the tip of the person's nose, so that visible features on profiles, for instance, should be in roughly similar locations to face-on. This 'window' could then be sub-sampled to a variety of resolutions for testing, ranging from 12×12 to 100×100. It was hoped that this would contain enough of the individual's facial features without too much distracting data, such as hair, background, etc.

The grey-levels were then compressed to remove unnecessary or superfluous details in the face (see Fig. 2). This left a narrow mid-range, where the majority of variation in facial details occured. This was found to be necessary in order

Figure 2: Example 25×25 subsampled face data

Figure 3: Example 21×21 convolved, binarised face data

to avoid the creation of confusing patterns, such as when changes in grey-level intensity in irrelevant regions such as hair highlights and background details became as prominent as the facial details such as eyes and mouth.

A 'difference of Gaussians' (DoG) convolution was then performed on the image to combine edge boundary detection with Gaussian smoothing. This process finds *zero-crossings*, originally described in Marr (1982), which are locations where the second derivative of the intensity values in the image undergo a sign change, and are especially useful for object segmentation. The idea of DoG-style *valley-detecting* convolution, where the 'width' scale is adjusted to be sensitive to face-sized features, is discussed in Bruce (1988) as being particularly optimal for face processing.

The final step was threshold binarisation (see Fig. 3), a technique commonly used after convolution to aid viewing of zero-crossings information.

These processing techniques can be seen to act together to reduce the dimensionality of the data in order that a minimal set (hopefully containing most of the significant information) is passed on to the network. This is needed because training times for neural networks are extremely sensitive to the scale of the data, and the quality of generalisation can also be adversely affected by extraneous data.

Since this is a classification task, there is an assumption that the image represents a human face, so no 'non-face' response is possible. However, the statistical nature of the output could lead to a 'confidence measure' from the level of output. Currently, classification of input by the network is simply given by the output node with the highest value.

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 10×10 data to several minutes to an hour for 90×90.

3.3 Shift and Scale Invariance Data Sets

Two data sets were created from the original data to test the RBF network's generalisation abilities: A scale invariance data set was produced with five copies of each image: one at the standard sampling 'window' size, two smaller (-12.5% and -25% of the standard surface area), and two larger ($+12.5\%$ and $+25\%$). An offset invariance data set was produced also with five copies of each image: one at the standard sampling 'window' position, and four others at the corners of a box centred on the standard position such as all x,y positions were $\pm$ 10 pixels from the centre.

4 RBF Network Results

4.1 Type I Data with Varying σ Values

Learning rate: 0.99, error limit: 0.005.

These tests were carried out with 20 images from 2 faces to check the influence of a fixed global (mean) σ value of the hidden units in the network on the success of identifying test patterns, to check its' influence on generalisation. The intuitive sense was that a higher σ would be better at generalisation, but might become confused with similar patterns of different classes.

It was observed that, as the σ value grew larger, the success rate could sometimes fall as it 'levelled-out' (see Fig. 4). This indicated that not only are large σ values inefficient (needing long training times), but they do not necessarily give the best results. This may be due to insufficiently localised receptive fields.

These tests showed that there is clear relationship between σ for the hidden unit response and the classification performance of the network. To automate the calculation of an effective σ value, the mean Euclidean distance between each hidden unit α and all others was used in the global form of the formula from Stokbro et al. (1990):

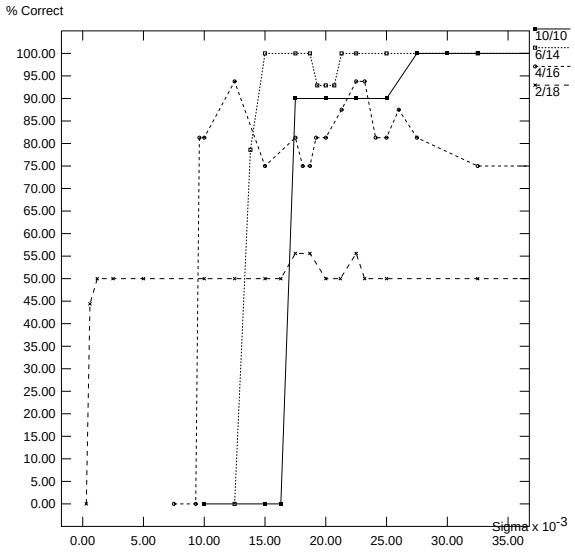

Figure 4: Comparison of generalisation to σ with 90×90 face samples

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

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 error measure value. If training is continued to minimize the error value to what would normally be considered as reasonable amount (< 0.1), the success rate drops to 50%. Since there are less patterns to be learnt here, it seems that the network becomes very prone to over-training.

In general, however, the lower the error limit, the better the training.

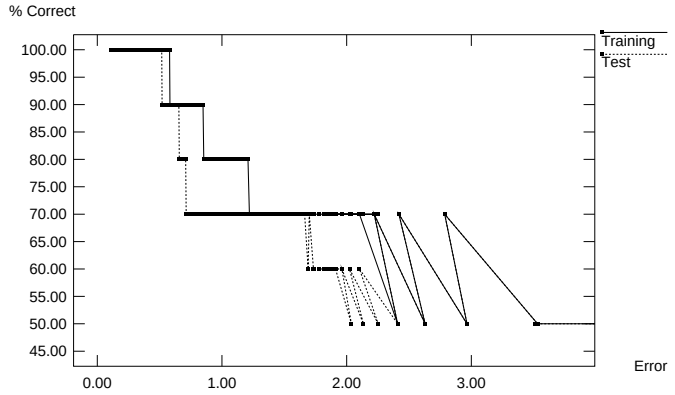

Figure 5: Comparison of error limit to recall (training images) and generalisation (test images) for 10/10 training with 44×44 face samples

4.3 Type I Data with Fixed Local σ Values

Learning rate: 0.99, error Limit: 0.005.

These tests were to check the success of local automated σ values.

Test 1: 90×90 face samples

Training/Test Patterns	Mean σ Set	Epochs	% Correct
10/10	0.05	3749	100
6/14	0.05	1507	100
4/16	0.06	866	75
2/18	0.05	55	50

Test 2: 44×44 face samples

Training/Test Patterns	Mean σ Set	Epochs	% Correct
10/10	0.10	1989	100
6/14	0.10	841	100
4/16	0.11	670	75
2/18	0.10	88	50

Test 3: 21×21 face samples

Training/Test Patterns	Mean σ Set	Epochs	% Correct
10/10	0.21	4942	100
6/14	0.26	3179	93
4/16	0.27	1664	75
2/18	0.24	57	50

Test 4: 10×10 face samples

Training/Test Patterns	Mean σ Set	Epochs	% Correct
10/10	0.54	11271	100
6/14	0.55	5290	86
4/16	0.57	2450	88
2/18	0.55	55	50

One test was made to check the performance of one global automated σ value, taken as the mean of individual values for the hidden units. This showed the same success rates, but with slower convergence, showing the advantage of local σ values.

These tests showed that an automated σ value, especially when local, was effective for training the network, as these results are equal in all cases to the maximum success rates recorded in the tests with a varying σ value.

The small difference in performance between the resolutions, and the computational expense of the larger patterns suggested that the 21×21 or 44×44 data sets should be used for further investigation.

4.4 Type I Data - Invariance Testing

Learning rate: 0.95, error limit: 0.05, random selection, run 100 times.

These tests are the most important yet, in that they checked the resilience of the network to offset and size variance of the data patterns. The error limit value was raised to 0.05 in the light of the results (shown in Fig. 5), which showed that little improvement in performance was gained with error limit values < 0.1 or so. This higher value showed significant reductions in processing times, which became more important as the amount of data increased.

Test 5: 21×21 face samples, offset variance

Training/Test Patterns	Min/Max Epochs	Min/Max % Correct
40/60	32868/109319	78/97
20/80	2684/4898	71/91
10/90	658/1376	64/80

Test 6: 21×21 face samples, size variance

Training/Test Patterns	Min/Max Epochs	Min/Max % Correct
40/60	6642/20120	92/93
20/80	1378/1820	79/91
10/90	648/1077	72/91

These tests show that the RBF network can work extremely well with even quite high degrees of variation in the image data, showing a performance equal (in some cases, higher) than for the non-variant data.

4.5 Type I Data - Clustering Test

Moody & Darken (1989) proposed the use of the *K-means* clustering algorithm to reduce the number of hidden units, however Musavi et al. (1992) described a simpler procedure, which was followed with slight adaptations. The distance d_{opp} of the nearest cluster of different class is compared to potential clusters made up of random combinations of existing clusters of the same class. If $d_{opp} > \alpha R$, where α is a constant ($1 \leq \alpha \leq 3$), the new cluster is accepted. High values of α encourage clustering and low values discourage.

This procedure was adapted by assuming that $R = 1$ with unit-normalised vectors, and by doing an exhaustive search of all potential clusters. This search iteratively selects the cluster with the largest d_{opp} value (assuming that it is greater than α). The disadvantage with this is that it tends to be very slow, and may become unusable in larger applications as the number of clusters increases.

The actual results with 21×21 samples, offset variance and 20/80 training showed *slower* convergence as the hidden units were clustered, with unpredictably fluctuating success rates.

This seems to indicate that clustering is not required and that the hidden layer is already at its optimal size. Note that Stokbro et al. (1990) recommends adding N hidden units to *each* original hidden unit (where N is the input vector dimensionality) for efficient interpolation. As this would have led to a hidden layer of hundreds or thousands of units in our case, we would then have needed clustering.

4.6 Type II Data Tests

Learning rate: 0.99, error limit: 0.05.

Due to the increased amount of data (100 images) to process for the ten people, these tests concentrated on the 21×21 face samples only. Random selections

were taken from each group to allow a useful spread of data.

Test 7: No variation, σ set individually

Training/Test Patterns	Min/Max Epochs	Min/Max % Correct
50/50	11161/25138	72/78
20/80	2102/2532	58/64
10/90	254/274	31/43

Test 8: Offset variation, σ set individually

Training/Test Patterns	Min/Max Epochs	Min/Max % Correct
100/400	325063/732436	32/37
60/440	56532/373322	28/34
50/450	14303/25788	28/31
40/460	12880/15311	25/27
20/480	2114/2886	19/23

Test 9: Size variation, σ set individually

Training/Test Patterns	Min/Max Epochs	Min/Max % Correct
100/400	413613/912466	30/38
60/440	58673/119653	27/35
50/450	12499/25832	27/31
40/460	12380/14150	25/29
20/480	1748/2931	18/23

4.7 Remarks

It can be seen that the RBF network performs significantly better than a purely random choice of the classes, which would give 50% for the Type I (2 face) data and 10% for the Type II (10 face) data. As with the Type I tests, the variant data gave similar performance to the non-variant where the train/test ratios were the same.

5 BP Network Results

A standard two-layer BP network was used for comparison. Random 'noise', added to the patterns when presented for training (not recall), helped with generalisation and the speed and success of convergence. This 'noise' can be

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 success rate of 50% in one training set but a more sophisticated metric relying on the differences between output values could be used to separate those with below average differences and boost the success rate to 80% of those that remain. How useful this would be in practise remains to be determined but it shows promise for situations in which further evidence (another close-up) can be gathered to get confident judgements of identity.

Another avenue of further investigation is the use of these techniques in more real-life applications. Here it will be necessary to determine ways of becoming robust to lighting variations, facial expressions, differences in appearance due to make-up, hairstyles etc. In addition, it will be useful to show that the RBF techniques can cope with rotations of view about the x and z axis as well as the y axis as tested so far. The simplification of the test conditions used so far allowed us to test the factors that seemed to be the most important (human heads move primarily around the y axis in upright cannonical views). The results are promising but more work is required to develop a useful system for face recognition.

Acknowledgments

Many thanks to the following who allowed their faces to be used for test data: Rachel Bundy, Julie Coultas, Erica Morris, David Nicholson, Alex Payne, Rafael Perez, Kathy Scott, Ben Shanks, Jabe Wilson.

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.

Bruce, V. (1988), *Recognising Faces*, Lawrence Erlbaum Associates.

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**, 302–309.

Ellis, H. D. & Young, A. W. (1989), Are faces special?, *in* A. W. Young & H. D. Ellis, eds, 'Handbook of Research on Face Processing', North-Holland.

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

Hay, D. C. & Young, A. (1982), The human face, *in* H. D. Ellis, ed., 'Normality and Pathology in Cognitive Functions', Academic Press.

Hay, D. C., Young, A. & Ellis, A. W. (1991), 'Routes through the face recognition system', *Quarterly Journal of Experimental Psychology: Human Experimental Psychology* **43**, 761–791.

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

Howell, A. J. (1993), Classification of human faces using neural networks, Master's thesis, University of Sussex. (Unpublished).

Marr, D. (1982), *Vision*, Freeman.

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

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

Musavi, M. T., Ahmad, W., Chan, K. H., Faris, K. B. & Hummels, D. M. (1992), 'On the training of radial basis function classifiers', *Neural Networks* **5**, 595–603.

Poggio, T. & Girosi, F. (1990), 'Regularization algorithms for learning that are equivalent to multilayer networks', *Science* **247**, 978–982.

Stokbro, K., Umberger, D. K. & Hertz, J. A. (1990), 'Exploiting neurons with localized receptive fields to learn chaos', *Complex Systems* **4**, 603–622.

Turk, M. & Pentland, A. (1991), 'Eigenfaces for recognition', *Journal of Cognitive Neuroscience* **3**(1), 71–86.