Proc. International Conference on Automatic Face & Gesture Recognition, Killington, VT, pp. 224–229, 1996

Towards Unconstrained Face Recognition from Image Sequences *

A. Jonathan Howell 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 presents experiments using Radial Basis Function (RBF) networks to tackle the unconstrained face recognition problem using low resolution video information. Input representations that mimic the effects of receptive field functions found at various stages of the human vision system were used with RBF networks that learnt to classify and generalise over different views of each person to be recognised. In particular, Difference of Gaussian (DoG) filtering and Gabor wavelet analysis are compared for face recognition from an image sequence. RBF techniques are shown to provide excellent levels of performance where the view varies and we discuss how to relax constraints on data capture and improve preprocessing to obtain an effective scheme for real-time, unconstrained face recognition.

1. Introduction

A face recognition system must be robust with respect to the immense variability of the human face and generalise over a wide range of conditions to capture the essential similarities for each individual. It is only recently that work on biologically-motivated, statistical approaches to face recognition has begun to deliver real solutions [1, 3, 13, 16]. One of the main problems that these approaches tackle is dimensionality reduction to remove much of the redundant information in the original images. There are many possibilities for such representations of the data, including principal component analysis, Gabor filters and various isodensity map or feature extraction schemes. In particular, it seems that appropriate preprocessing of input representations for a face recognition scheme can overcome the problems of lighting variation and multiple scales. Other sources of variation such as face orientation, expression, occlusion etc. still remain.

In our work [7, 8, 9, 10] we use an adaptive learning component based on RBF networks to tackle the face recognition problem. We want our face recognition scheme to generalise over a wide range of conditions to capture the essential similarities of a given face. The RBF network has been identified as valuable model by a wide range of researchers [2, 6, 14, 18]. 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 good for practical vision applications as they provide a guaranteed, globally optimal solution via simple, linear optimisation. In this paper we contrast the use of DoG filters and Gabor wavelet analysis as input representations for our networks. More important, we also extend our work on static images to the time vaying case where we have to recognise an individual in an image sequence. In general, this involves a greater degree of variability in scale, shift, pose, and expression as we cannot pick and choose the views. We first consider training and testing on images from a seated subject where scale and shift are quite constrained. Then we go on to consider the case where the subject is free to walk about and is tracked (imperfectly). Training and testing in this unconstrained situation is much more problematic. We consider schemes to enforce temporal consistency and to use the confidence measure delivered by the RBF networks.

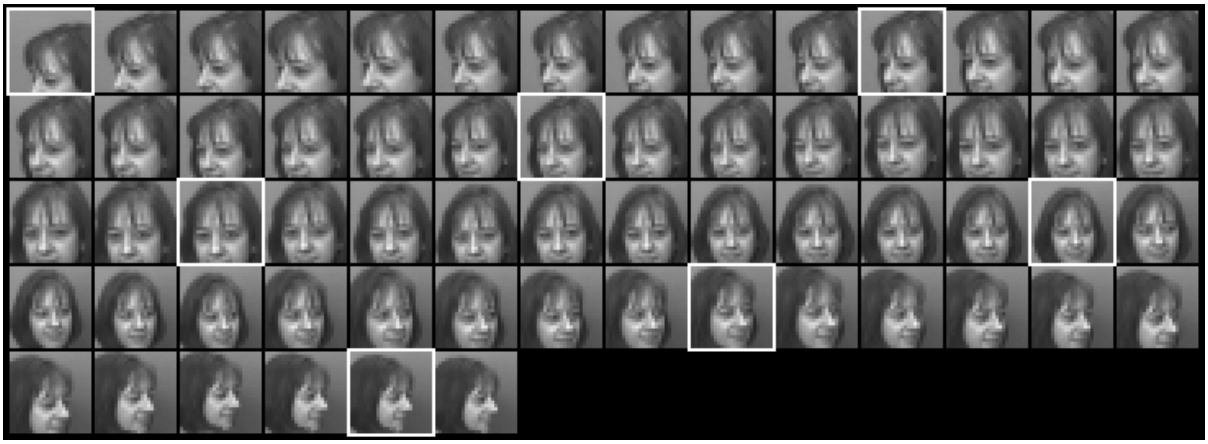

Figure 1. A complete Primary sequence for class *carla*, after segmentation but before pre-processing (boxes indicate frames used for training with selection interval of 10).

2. Description of Problem

Initial research often requires restrictions on the variability of test data in order that fundamental principles can be investigated in isolation. However, this means that such applications are far removed from real-world environments, where data is noisy and unpredictable. Besides a theoretical desire to remove constraints, there is a real commercial demand for a system that can rapidly identify a person from a small group of users.

In this paper, we address the problem of recognising individuals from a small group (less than 100) in real-time from low-resolution video image sequences. Face recognition is a computationally expensive process and to obtain such real-time performance requires certain trade-offs. A police record application would require access to enormously large amounts of data but accuracy would have priority over speed, as instantaneous recognition would not be the primary factor. To cope with hundreds of thousands of individuals, views may be limited to face-on or profile only with the face at a specific region of the image, allowing precise pin-pointing and measurement of feature points. We are considering a less constrained environment, where people can move around freely, and so we need to recognise the person from the full range of views where the face is visible.

In addition, the example police application would require extremely low error rate, and only very few (maybe even just one) image of each individual would be available. We have opted for a lower accuracy method that is considerably faster, and provides a reasonable discarding of low-confidence output. The use of image sequences means we have an enormous abundance of data, and if the current image is ambiguous, it can be discarded and the next considered (see Figure 1). The temporal coherence of human faces allows the matching of series of frames linked by movement information with the use of 'time windows' to combine information from several frames.

3. The RBF Network Model

The RBF network is a two-layer, hybrid learning network [14, 15], 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. Each function has an associated centre width value which defines the nature and scope of the unit's receptive field response, giving 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, as very low (or no) output will occur if the pattern is more than slightly different to those trained, allowing the removal of outliers.

The weights can be adjusted using the Widrow-Hoff [19] delta learning rule, however, the single layer of linear output units permits a matrix pseudo-inverse method [18] for their exact calculation. The latter approach allows training of the network in a small fraction of a second. In the test phase, 500 images are processed in around two seconds, giving a single classification in around 4 ms, which is already adequate for real-time sequences.

4. Specification for Image Sequences

The image sequences used in the tests reported here are the result of collaboration with Stephen McKenna and

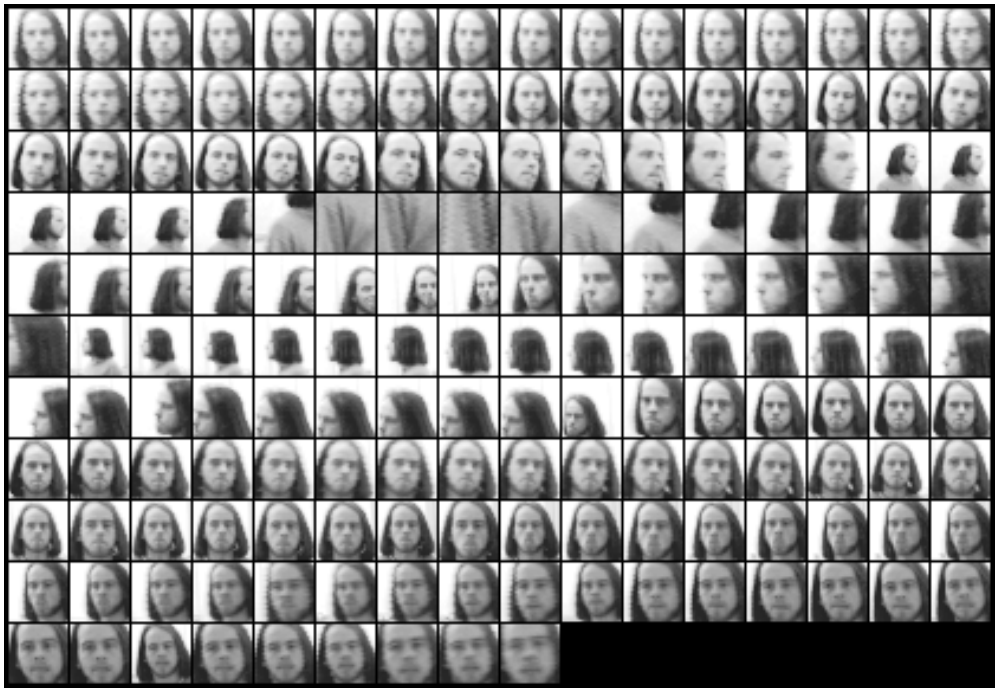

Figure 2. A complete Secondary sequence for class *steve*, after segmentation but before pre-processing. As only front-view face detection has been implemented at this stage, some non-face frames are included and profile views, although segmented, are incorrectly scaled.

Shaogang Gong at Queen Mary and Westfield College (QMW), University of London, who are researching real-time face detection and tracking [12]. We have devised two types of sequences to simulate a (fairly) unconstrained environment, termed 'Primary' and 'Secondary'.

The intention is to train the network with a controlled set of data – the Primary image sequences – known to include the types of variability which we want our trained system to be tolarant to (thus including 180° range of pose angles but a blank background), and to test on totally unrelated data – the Secondary image sequences. This total separation of training and test data is to allay any fears that the system is using spurious environmental details, such as lighting or background features, to classify individuals. This problem can always appear in databases where test and training data are collected at the same time and manner or arbitrarily selected from a central database.

The **Primary** image sequences are intended to provide suitable data to train the system for an on-line source of test images. They consist of a person moving from one profile view to the other whilst sitting on a chair (to limit body movement) against a plain, mid-grey background (to limit background effects). Eight Primary sequences have been collected so far, each featuring a different person. They range in length from 62 frames to 94 frames, 554 images in total.

The **Secondary** image sequences are intended to simulate an on-line source of test images, and are much more variable than the Primary image sequences to simulate tracking in an unconstrained envirionment. They will consist of fairly long sequences of one person moving around a room, allowed to move from side to side and stop and start movement against a cluttered, changing background. Only one preliminary Secondary sequence has been collected so far; this has 169 frames.

A typical sequence of images from a motion-based head tracker such as Figure 1 illustrates that perfect registration of the head and face can never be guaranteed, except by manual methods. Future developments of the face detection scheme can be expected to discard any non-face frames, improving recognition, whilst maintaining temporal continuity.

5. Pre-processing of Segmented Data

Two main techniques are used for the preprocessing of the images: Difference of Gaussian (DoG) filtering and Gabor wavelet analysis at a range of scales. One way of thinking about these input representations and mapping them onto our RBF networks is to use the analogy with visual

neurons. The receptive field of such a neuron is the area of the visual field (image) where the stimulus can influence its response. For the different classes of these neurons, a receptive field function $f(x, y)$ can be defined. For example, retinal ganglion cells and lateral geniculate cells early in the visual processing have receptive fields which can be implemented as DoG filters [11]. Later, the receptive fields of the simple cells in the primary visual cortex are oriented and have characteristic spatial frequencies. Daugman [5] proposed that these could be modelled as complex 2-D Gabor filters. Petkov *et al.* [17] successfully implemented a face recognition scheme based on Gabor wavelet input representations to imitate the human vision system. Our earlier studies [8] showed that these later stages of processing make information more explicit for our face recognition task than the earlier DoG filters.

In contrast to more deterministic methods using warping based on registration of features, eg [4], our approach uses simpler preprocessing, but learns to discriminate using the RBF networks to overcome occlusion arising out of head rotation.

The experiments presented here concentrate on two specific applications of these techniques:

DoG convolution with a scale factor of 0.4, with a reduced range of grey-levels, with thresholding to give *zero-crossings* information. A 25×25 image gave 441 samples per image.

Gabor 'A3' sampling (described in [8]), with a full range of grey-levels. Four non-overlapping scales were used with three orientations including sine and cosine components. A 25×25 image gave 510 coefficients per image.

Images at set intervals were extracted from the Primary sequences in order to train the network, all the others were used for testing. This Train/Test ratio is recorded below.

6. Results

To test the ability of the RBF network to classify test images after training with the Primary sequences, experiments were done initially by dividing the Primary sequences into training and test groups:

(a)

Interval	Train/ Test	Initial %	% Discarded	% After Discard
2	278/276	96	12	98
5	114/440	88	30	99
10	60/494	75	50	90
20	33/521	58	68	90
30	24/530	48	81	93
50	16/538	40	81	86

(b)

Interval	Train/ Test	Initial %	% Discarded	% After Discard
2	278/276	99	2	100
5	114/440	98	7	100
10	60/494	95	16	98
20	33/521	87	35	94
30	24/530	73	55	94
50	16/538	67	62	94

Table 1: Effect of selection interval and pre-processing methods using a standard RBF Network trained and tested on the eight Primary sequences: (a) DoG (b) Gabor.

It can be seen that although the initial results tail off as the sampling interval increases, the confidence discard allows the maintenance of a high standard of performance. In particular, that the network can still recognise 94% of the images with a pose angle range of 180° , having been trained with only two for each class (at a sampling interval of 50), is a great achievement.

The Secondary sequences are still in development, but encouraging results (see Table 2) have been collected from preliminary experiments with a provisional sequence.

(a)

Interval	Train/ Test	Initial %	% Discarded	% After Discard
2	278/169	43	69	42
5	114/169	32	76	19
10	60/169	44	75	35
20	33/169	23	76	21

(b)

Interval	Train/ Test	Initial %	% Discarded	% After Discard
2	278/169	61	41	77
5	114/169	56	45	77
10	60/169	60	43	81
20	33/169	54	42	66

Table 2: Effect of pre-processing methods using a standard RBF Network trained on eight Primary sequences and tested with a separate Secondary sequence: (a) DoG (b) Gabor.

7. Temporal Integration

An on-line source of data cannot be evaluated like a conventional database, as not all of the data is yet present, nor is it known who will be present in the environment or for how long. If an assumption of *temporal coherence* is made, ie that one person will not transform into another instantly, high-confidence information in a 'time window'

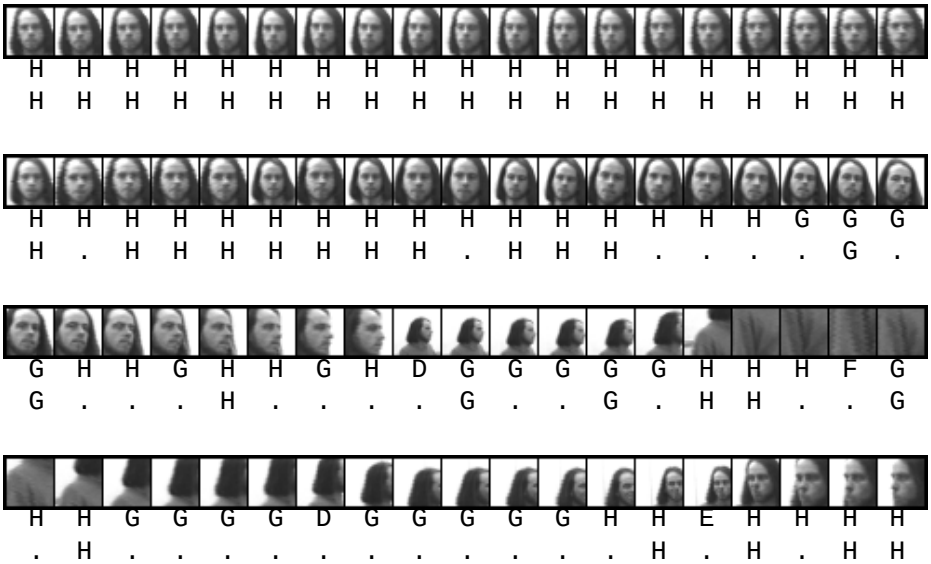

Figure 3. Typical output for a section of the Gabor preprocessed Secondary sequence containing class *steve* **(H). Top row of letters shows initial output, lower row output after discard of low-confidence values ('.' indicating such a discarded value).**

can be utilised in periods of low-confidence output to lend support to the current output (assuming it is the same classification, though conflicting output could also be useful). In this envirionment, a running total of output values for a time window can be kept, and a expression for the individual currently in view given. For this to work, fairly low (above random) success rates will suffice.

To illustrate this technique, consider Figure 3, which shows a section of the test Secondary sequence. Table 3 shows how use of time windows to re-assess the output value can affect performance. Periods of correct output followed by random values can be interpreted as all correct, using the last stable output as a type of memory. It can be seen that such techniques which take advantage of temporal coherence can be used to improve performance.

Time Window	Initial %	% After Discard
1	66	86
3	72	86
5	68	89
10	68	100

Table 3: Results using varying time windows in ouput for the sequence shown in Figure 3.

8. Performance with More Classes

It must be emphasised that this research is at a preliminary stage but that the technique shows promise for scaling up to large databases. Although only a few individuals are shown in our image sequences, this type of network has been shown to work well with larger numbers of classes. For example, the Olivetti Research Laboratory database of faces[1] with 400 images of 40 people can be distinguished with a high level of performance (see Table 4).

Images per Person	Train/ Test	Initial %	% Discarded	% After Discard
5	200/200	84	39	95
4	160/240	80	43	95
3	120/280	72	52	91
2	80/320	64	60	87
1	40/360	46	70	81

Table 4: Results for ORL Face Database, using Gabor preprocessing, averaged over two different selections.

9. Conclusion/Future Work

Several points can seen from the results:

[1]available via *ftp* for comparative tests, further information is at: `http://www.cam-orl.co.uk/facedatabase.html`

1. The RBF network is shown to generalise well from samples in classifying faces (3-D complex shapes) in real-time sequences.

2. Gabor preprocessing is shown to give a more generally useful input representation than the DoG preprocessing, especially for the more difficult Secondary sequence.

3. The confidence measure used in discarding uncertain classifications is shown to be important for handling sequences especially where a small training set is used.

In conclusion, the locally-tuned linear Radial Basis Function (RBF) networks showed excellent performance in the simpler face recognition task when trained and tested on images from Primary sequences. This is a promising result for the RBF techniques considering the high degree of variability due to the varying views (mostly rotations) of a person's face in these data sets. The results so far from the Secondary sequences also show considerable promise, especially with the additional use of temporal coherence to improve performance. In the these sequences, the face detection scheme [12] currently selects and rescales faces in near face-on views but does not discard the others. It is expected that further development of this scheme will allow improvements in the reliable and consistent labelling of faces in unconstrained image sequences. It is clear that the ability of the RBF networks to give a measure of confidence, which allows temporal integration over image frames where the visual evidence is poor, is essential for this development.

Work is progressing together with colleagues at QMW in refining the face detection scheme and automated on-line learning of new classes of individual. The next stage of development will integrate this refined on-line face detection and localisation with the trained RBF networks to cope with real-time image sequences including the usual variations in illumination as well as position, scale, view and facial expression. It is clear from the work of [2] and others that using statistically based techniques is the key to good performance. The RBF techniques are mathematically well-founded, which gives a clear advantage in engineering a solution to our application problems.

References

[1] D. Beymer and T. Poggio. Face recognition from one example view. In *Proc. Int. Conference on Computer Vision*, pp. 500–507, Cambridge, MA, 1995.

[2] C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.

[3] R. Brunelli and T. Poggio. Face recognition: Features versus templates. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 15:1042–1052, 1993.

[4] I. Craw, N. Costen, T. Kato, G. Robertson, and S. Akamatsu. Automatic face recognition: combining configuration and texture. In *Proc. Int. Workshop on Face and Gesture Recognition*, pp. 53–58, Zurich, Switzerland, 1995.

[5] J. G. Daugman. Complete discrete 2-D gabor transforms by neural networks for image analysis and compression. *IEEE Trans. Acoustics, Speech, and Signal Processing*, 36(7):1169–1179, 1988.

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

[7] A. J. Howell and H. Buxton. Invariance in radial basis function neural networks in human face classification. *Neural Processing Letters*, 2(3):26–30, 1995.

[8] A. J. Howell and H. Buxton. Receptive field functions for face recognition. In *Proc. 2nd Int. Workshop Parallel Modelling of Neural Operators for Pattern Recognition*, Faro, Portugal, 1995.

[9] A. J. Howell and H. Buxton. A scaleable approach to face identification. In *Proc. Int. Conference on Artificial Neural Networks*, volume 2, pp. 257–262, Paris, France, 1995.

[10] A. J. Howell and H. Buxton. Face recognition using radial basis function neural networks. In *Proc. British Machine Vision Conference*, Edinburgh, 1996.

[11] D. Marr and E. Hildreth. Theory of edge detection. *Proc. R. Soc. London*, B207:187–217, 1980.

[12] S. McKenna and S. Gong. Combined motion and model-based face tracking. In *Proc. British Machine Vision Conference*, Edinburgh, 1996.

[13] B. Moghaddam and A. Pentland. Probabilistic visual learning for object detection. In *Proc. Int. Conference on Computer Vision*, Cambridge, MA, 1995.

[14] J. Moody and C. Darken. Learning with localized receptive fields. In D. Touretzky, G. Hinton, and T. Sejnowski, editors, *Proc. 1988 Connectionist Models Summer School*, pp. 133–143, 1988.

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

[16] A. Pentland, B. Moghaddam, and T. Starner. View-based and modular eigenspaces for face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 84–91, 1994.

[17] N. Petkov, P. Kruizinga, and T. Lourens. Biologically motivated approach to face recognition. In *Proc. Int. Workshop on Artificial Neural Networks*, pp. 68–77, 1993.

[18] T. Poggio and F. Girosi. Networks for approximation and learning. In *Proc. IEEE*, volume 78, pp. 1481–1497, 1990.

[19] B. Widrow and M. E. Hoff. Adaptive switching circuits. In *1960 IRE WESCON Convention Record*, volume 4, pp. 96–104. IRE, New York, 1960.