International Journal of Neural Systems, Vol. 15, Nos. 1&2 (2005) 1–8

World Scientific
www.worldscientific.com

THE ROLE OF GLOBAL AND FEATURE BASED INFORMATION IN GENDER CLASSIFICATION OF FACES: A COMPARISON OF HUMAN PERFORMANCE AND COMPUTATIONAL MODELS

SAMARASENA BUCHALA*, NEIL DAVEY†, RAY J. FRANK‡ and MARTIN LOOMES§
School of Computer Science, University of Hertfordshire,
College Lane Hatfield, Herts. AL10 9AB, UK
**S.Buchala@herts.ac.uk*
†*N.Davey@herts.ac.uk*
‡*R.J.Frank@herts.ac.uk*
§*M.J.Loomes@herts.ac.uk*

TIM M.GALE
Department of Psychiatry, QEII Hospital,
Welwyn Garden City, Herts., AL7 4HQ, UK
T.Gale@herts.ac.uk

Most computational models for gender classification use global information (the full face image) giving equal weight to the whole face area irrespective of the importance of the internal features. Here, we use a global and feature based representation of face images that includes both global and featural information. We use dimensionality reduction techniques and a support vector machine classifier and show that this method performs better than either global or feature based representations alone. We also present results of human subjects performance on gender classification task and evaluate how the different dimensionality reduction techniques compare with human subjects performance. The results support the psychological plausibility of the global and feature based representation.

Keywords:

1. Introduction

Most computational models of gender classification use whole face images, giving equal weight to all areas of the face, irrespective of the importance of internal facial features. In this paper we evaluate the importance of global and local information in a series of gender recognition experiments. Global processing of faces is assumed to encode coarse information like shape and configuration of internal features, while featural processing utilises more detailed representations of facial features, such as the eyes and mouth. In psychological terms, the latter implies an attentional component whereby salient features are processed in more detail than the coarse image. In this study we use these two kinds of representation. Since face image data have a very high dimensionality, we apply dimensionality reduction techniques on the data before applying a Support Vector Machine (SVM) to classify gender. For comparison we use different dimensionality reduction techniques, such as Principal Component Analysis (PCA), Curvilinear Component Analysis (CCA), and Self Organising Maps (SOM). Finally, we present results of human subjects performance on gender classification task and evaluate how the different dimensionality reduction techniques compare

with human subjects performance. The main findings of this study are as follows.

1. Gender classification of the global and featural model is significantly better than either global (full face) or featural models (eyes and mouth).
2. All three dimensionality reduction techniques produced high classification rates, with PCA performing slightly better than CCA and SOM. However CCA, a nonlinear method, needed far fewer variables compared to PCA.
3. Experiments with human subjects showed impressive levels of gender recognition accuracy from representations of single facial features (i.e., eyes and mouths). This underscores the importance of these specific features and supports the psychological plausibility of the global and feature based model discussed in this paper. Moreover, there was some association between the errors made by the models and those made by human observers.

The remainder of the paper is organised as follows. A brief introduction of the different dimensionality reduction approaches used in this study is presented in the next section. Related work is discussed in Sec. 3. Section 4 presents the global and feature based method used for this study. Sections 5 and 6 present the computational and human experimental results. We conclude with some discussion of the results in Sec. 7.

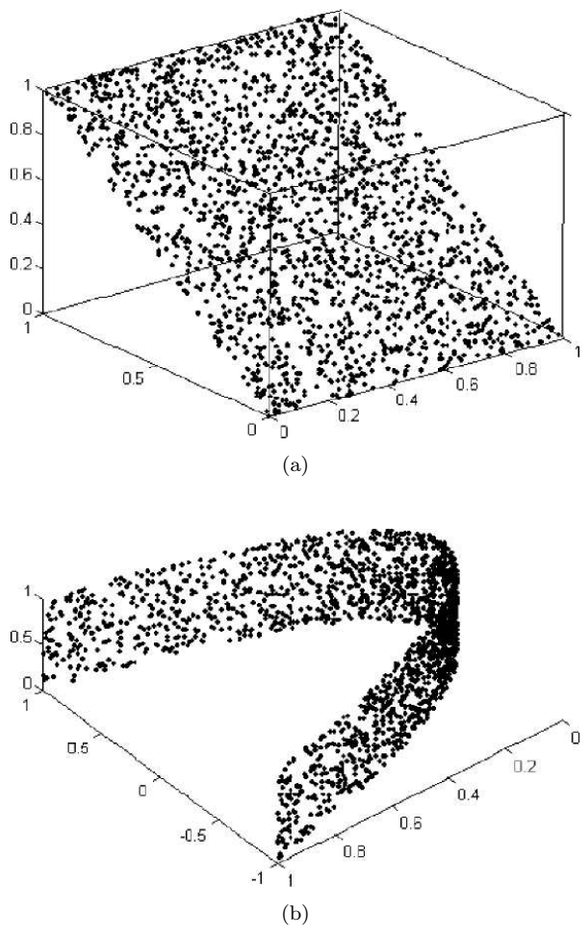

Fig. 1. Linear and nonlinear data.

2. Dimensionality Reduction

High dimensional data usually contain redundancies and may have many irrelevant variables. Classifiers like neural networks may need huge networks, with many free parameters, to cover the high dimensional data. Networks, on such datasets, even if successfuly trained, often perform badly on their test sets. This bad generalization may be due to the large number of free parameters representing irrelevant information. To learn relevant information from such datasets, a large number of datapoints would be needed, which is often impractical, and the training time needed for learning also increases to a great extent. This problem with high dimensional data is often referred in the literature as the "curse of dimensionality".[1]

Due to correlations among the data, linear and nonlinear, a D dimensional data may actually lie on a d dimensional manifold ($D > d$) and the true dimension or the Intrinsic Dimension of such data is said to be d. For example a plane embedded in a three-dimensional space, as shown in Image (a) in Fig. 1 has an Intrinsic Dimension value of 2 as data on two axes are linearly dependent. Image (b) in Fig 1 shows the well known three dimensional horseshoe data distribution. However any point in the data can be defined by a linear axis and a curvilinear axis, indicating that its Intrinsic Dimension value is 2. The problems, related to high dimensionality, can be circumvented by accounting the correlations among the data and reducing the data to its Intrinsic Dimension. There is also evidence that redundancy reduction is an important part of sensory processing in human brain.[2]

2.1. *Principal Component Analysis*

Principal Component Analysis (PCA)[3] is a popular dimensionality reduction technique that linearly transforms a D dimensional data to a d dimensional data, without significant loss of information, where $d \leq D$. PCA finds principal component axes in the data cloud so that data dimension can be reduced by projecting the data onto these axes. The first principal component would be in a direction, such that it accounts for the maximum variance of the data. The second principal component lies in a direction normal to the first principal component, and accounts for as much of the remaining variance as possible. The third principal component would be in a direction normal to the first two and so on.

2.2. *Curvilinear Component Analysis*

Curvilinear Component Analysis (CCA)[4] is a recent technique, which has the ability to account for strong nonlinear correlations among the data. The idea of CCA is to preserve distances in the input and output spaces; all the possible distances between points in the input space should match the respective distances in the output space. However, preservation of larger distances may not be possible in the case of nonlinear data, as a global unfolding of the manifold is required to reduce the dimension. In this case, it is important that at least local (smaller) distances should be preserved. For this, CCA uses a neighborhood function which ensures the condition of distance matching is satisfied for smaller distances while it is relaxed for larger distances. Preservation of smaller distances (local mapping) may then lead to the stretching of larger distances (global unfolding). The projection layer of CCA minimizes an error function which is given as

$$E = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(d_{i,j}^{X} - d_{i,j}^{Y}\right)^2 F_\lambda\left(d_{i,j}^{Y}\right) \quad \forall\, j \neq i \quad (1)$$

where $d_{i,j}^{X}$ and $d_{i,j}^{Y}$ are the Euclidean distances between points i and j in the input space X and output space Y respectively. $F_\lambda(d_{i,j}^{Y})$ is the neighborhood function, selected such that it favors smaller distances over larger ones. Minimizing the error function with respect to the point Y_i in the output space by a normal stochastic gradient would give the following update rule.

$$\nabla Y_i = \alpha(t)\sum_{j=1}^{N}\left[2F_\lambda(d_{i,j}^{Y}) - (d_{i,j}^{X} - d_{i,j}^{Y})F'_\lambda(d_{i,j}^{Y})\right] \times \left[\frac{d_{i,j}^{X} - d_{i,j}^{Y}}{d_{i,j}^{Y}}\right](Y_i - Y_j) \quad \forall\, j \neq i \quad (2)$$

$\alpha(t)$ the learning rate, and the neighborhood function $F_\lambda(d_{i,j}^{Y})$ can be time varying.

The stochastic gradient update method of (2) can be conceived as selecting a point Yi in the output space, while the remaining points are pinned. The selected point is moved (updated) according to the average influence of all the pinned points. This method of updating has the following drawbacks.[4]

- The computational cost is of the order of $O(N^2)$ as all the possible $N(N-1)/2$ distances need to be calculated at each time step.
- The sum of all influences may lead to an averaging effect, which leads to a small update amount resulting in slow convergence.

For these reasons CCA uses a different update method, where the selected point is pinned while the remaining points are moved according to its influence. Then, by ignoring the derivative part of (2), the update rule of CCA can be written as:

$$\nabla Y_j = \alpha(t)F_\lambda\left(d_{i,j}^{Y}\right)\frac{d_{i,j}^{X} - d_{i,j}^{Y}}{d_{i,j}^{Y}}(Y_j - Y_i) \quad \forall\, j \neq i. \quad (3)$$

The algorithm for projection of the training data can be summarized as follows:

```
Calculate the Euclidean distances between
   all pairs of points in the input space.
Initialize the points in the output space
   randomly or using PCA.
Initialize epoch t = 0
For each epoch t,
Begin
   Calculate alpha(t) and lambda.
   For each point Yj in the output space,
   Begin
```

$$\nabla Y_j = \alpha(t)F_\lambda\left(d_{i,j}^{Y}\right)\frac{d_{i,j}^{X} - d_{i,j}^{Y}}{d_{i,j}^{Y}}(Y_j - Y_i) \quad \forall\, j \neq i$$

```
   End
   Increment t
End
```

Mapping of a new point (test data) from the input space X to the output space Y, in CCA, involves reducing the error function of (1) and is iterative in the same sense as the actual learning process. However, the update rule is the stochastic gradient of (2) without the derivative part. The algorithm for projecting a new point can be summarized as follows:

```
Calculate the Euclidean distances between
  the new test point and each training point.
Initialize the test point in the output
  space randomly or using PCA.

Initialize epoch t = 0.
For each epoch t,
Begin
   Calculate alpha(t) and lambda.
```

$$\nabla Y_i = \alpha(t) \sum_{j=1}^{N} F_\lambda(d_{i,j}^Y) \frac{d_{i,j}^X - d_{i,j}^Y}{d_{i,j}^Y} (Y_i - Y_j) \quad \forall\, j \neq i$$

```
   Increment t
End
```

We use the first few variables obtained by the PCA projection, for initialization of the points in the output space. This initialization, rather than a random one, induces some prior information about the submanifold of the data. The learning rate and the neighborhood width are calculated as an exponential decay.

2.3. *Self Organising Map*

Self Organising Map (SOM)[5] is a well-known nonlinear method that learns a mapping from a D dimensional input space X to a d dimensional output space Y by using principles of Vector Quantization and Topological Mapping.

3. Related Work

Issues in gender classification have stimulated a great deal of research by psychologists and computer scientists. While the research in Psychology[6–8] has largely been within the context of human visual processing, and identifying key featural differences in males and females, Computer Science research[9–12] has been geared more towards specific face identification. The computational models range from using pixel-based information to representations derived from geometric measurements. Studies also vary considerably in the size of training sets used and in the type of features present or absent (for example, some studies use hair information while others do not). Nevertheless, most models, and specifically those that are pixel-based, have used whole face images, where the salience of specific facial features is not captured. These can be termed as global models.

4. Face Representation

Hair, especially for females, forms a major part of a facial image and has a dominating affect on classification. Abdi *et al.*[8] reported gender classification accuracy of 80% for hairless faces against 91.8% for the same faces with hair information included. However, in our previous work,[13] classification rates on faces, with hair information removed was better than that on faces with hair information. The performance degradation on faces with hair information in our experiments was due to the variability of hairstyles in the dataset. Despite these disparate results, hair can certainly be an important visual cue for gender identification. The first image in Fig. 2 shows a pictorial view of the difference in means of female and male face images. The lighter the pixel luminance, the larger is the difference and the darker the luminance, the smaller is the difference between means. This pictorial view suggests that regions around the face outline, chin, mouth, and above the eyes carry discriminatory information. However, the region around the face outline, with much brighter luminance, carries greater discriminatory information. This region signifies the presence or absence of hair. The second and third images of Fig. 2 are the pictorial views of the standard deviations within

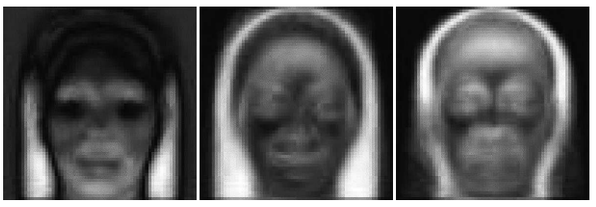

Fig. 2. Pictorial views of difference measures.

the female and male face images respectively. Again, the lighter the pixel luminance, the larger is the standard deviation. These images, however, indicate that the discriminatory information of the regions around neck and face outline is variable to a large extent in females and to a certain extent in males. From this simple analysis, it can be said that hair information is important. However, a psychologically plausible face-representation should overcome the problem of variable hairstyles.

In Fig. 2, the first image is the pictorial representation of the difference of the means, of female and male face images. The second image is the standard deviation within the female face images. The third image is the standard deviation within the male face images.

In Fig. 3, the three sub-images are obtained from the original 128×128 image. A 32×64 image pertaining to the eye region and a 32×64 image pertaining to the mouth region are extracted from the original image. The third sub-image is a 64×64 reduced resolution version of the original image.

In this study we use a global and feature based representation of face images which embodies both global and featural information. From a 128×128 face image, three sub-images are obtained as illustrated in Fig. 2. A 32×64 pixel strip pertaining to the eyes region, taking the midpoint between the two eyes as a reference point, and a 32×64 pixel strip pertaining to the mouth region, taking midpoint of the mouth as a reference point, are extracted from each face image. These sub-images represent salient featural information. The third sub-image is a 64×64 reduced resolution version of the original image and this represents global information. In this study, the quantity of pixel information is identical for featural and global representations. A similar type of face representation was also used by Luckman *et al.*[14] for their computational model of familiar face recognition.

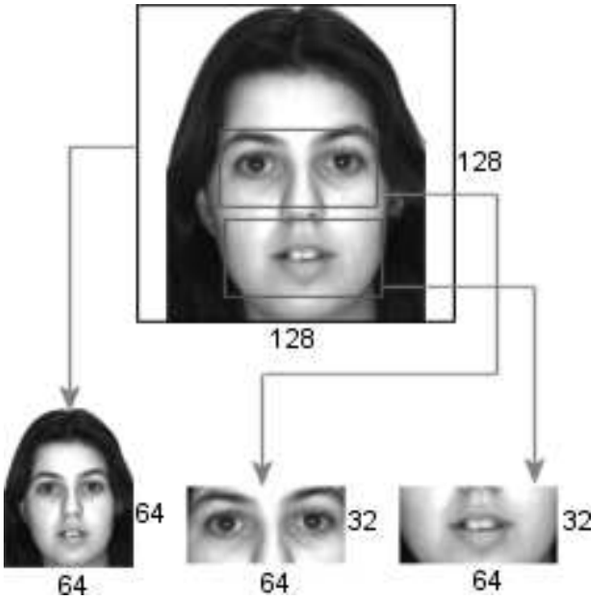

Fig. 3. Extracting sub-images.

5. Computational Experiments

Experiments are carried out using 400 frontal face (200 female and 200 male) greyscale images. The faces are from the following databases: FERET,[15] AR,[16] and BioId.[17] Three sub-images, as explained in the previous section, are extracted for each of the 400 faces. Histogram equalization is then applied on all three sub-images to normalize for different lighting conditions. We use five-fold cross validation, with 320 faces (160 females and 160 males) for each training set and 80 faces (40 females and 40 males) for each test set, and report average classification rates using an SVM[18] classifier, with RBF kernel. Before applying classification, dimensionality reduction techniques discussed in Sec. 3 are applied on the sub-image data. For PCA reduction we use the first few principal components, which account for 95% of the total variance of the data. Since CCA has the ability to reduce the dimensionality of strongly-nonlinear data, we use an *Intrinsic Dimension* estimation technique, the Correlation Dimension,[19] and reduce the data dimension to this *Intrinsic Dimension*. For SOM reduction, the subspace dimensionality is chosen as 64 (8×8 output grid) for the whole face and 36 (6×6 output grid) for eyes and mouth sub-images.

First we present classification results on the sub-images data. As shown in Table 1, all three sub-images produced high classification rates, indicating a surprisingly high amount of gender information in each of them. The figures in parentheses indicate the subspace dimensionality.

Classification is performed on the composite data, obtained by combining the data from the three sub-images. It can be seen from Table 2 that PCA performed marginally better than CCA and

Table 1. Average classification rates of the sub-images by an SVM. Figures in parentheses are the number of variables obtained after dimensionality reduction.

Feature	PCA	CCA	SOM
Eyes	85.5% (250)	82.75% (22)	80.25% (36)
Mouth	81.25% (253)	81.55% (22)	80.25% (36)
Full face	87.5% (256)	87.0% (26)	83.25% (64)

Table 2. Classification rates of the composite data and original image data by an SVM. Figures in parentheses are the number of variables obtained after dimensionality reduction.

Feature	PCA	CCA	SOM
Composite	92.25% (759)	91.5% (70)	89.75% (136)
Original full face	86.5% (283)	85.5% (36)	83.25% (81)

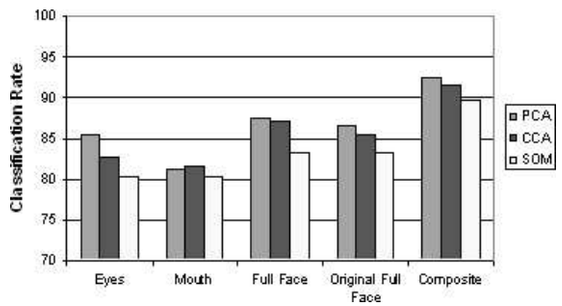

Fig. 4. Average classification rates on different features.

SOM. However, CCA uses far fewer variables (70) than PCA (759). For a comparison, we also report the classification rates of the data of the original 128×128 faces.

It can be seen from Table 2 that the composite data, which includes both global and featural information, performed significantly better than the global model. Figure 4 shows that the composite data outperformed all other data representations.

6. Human Experiments

6.1. *Eye images*

Mean performance accuracy for eye classification was 77.25% (standard deviation = 5.42%).

Table 3. A comparison of classification accuracy rates by human participants for mouth images classified incorrectly and correctly by the 3 computational models.

Model	Incorrect Items	Correct items	Human accuracy (incorrect)	Human accuracy (correct)
PCA	13	67	72.7%	78.4%
CCA	13	67	71.9%	78.6%
SOM	19	61	57.1%	83.8%

Chance performance on this task would be 50% so participants performed well above chance. There was no difference between male and female participants in terms of their accuracy. In the items analysis, gender recognition accuracy varied considerably across the 80 eye images (range 13–100% correct). Interestingly, there were very few sets of eyes that elicited chance levels of recognition performance. Rather, they tended to be correctly classified by the majority or incorrectly classified by the majority. A major focus of interest with this work is whether the classification errors of human participants are associated with those of the computational models (PCA, CCA, and SOM) under generalization. We subdivided the 80 eye images into 2 groups based on whether each model had classified the gender correctly. We then investigated whether those items that were erroneously classified by the model were less accurately classified by the 80 human participants. This analysis is summarized in the table below.

Although the accuracy of humans was always higher for items that had been correctly classified by the models, than that had been incorrectly classified, this difference was statistically significant only for the SOM ($p < 0.005$). Since the data were not normally distributed, differences were analyzed nonparametrically (with Mann-Whitneys U Test). It is notable that the SOM made more classification errors than the two other models and this may be why it predicts the human data more correctly. The other two models made few errors overall and hence the sample size is small.

6.2. *Mouth images*

Mean performance accuracy for gender classification of mouth images was 75.4% (standard deviation = 5.7%). The fact that, once again, participants scored

Table 4. A comparison of classification accuracy rates by human participants for mouth images classified incorrectly and correctly by the 3 computational models.

Model	Incorrect items	Correct items	Human accuracy (incorrect)	Human accuracy (correct)
PCA	15	65	57.0%	79.7%
CCA	20	60	54.6%	82.4%
SOM	21	59	59.2%	81.2%

well above chance level suggests that information useful for gender recognition can be derived from specific facial features, even when represented at a fairly low level of resolution. The overall accuracy rate of the models and human participants is very similar. As with the eye data, we compared human performance on those mouth images that the model had classified incorrectly and correctly. This data is presented in Table 4.

Similar to the results on the Eye sets, the mean accuracy of humans was always higher for items that had been correctly classified by the models, than that had been incorrectly classified. But, the differences were significant at $p < 0.001$ or less for all 3 methods, showing that those items which the models fail to categorize correctly are more likely to elicit gender recognition errors in humans.

7. Discussion and Conclusion

Hair, especially for females, forms a major part of the image and has a dominating affect on the classification. Many males with long hair and females with short hair were misclassified when the original full face images are used. The global and feature based model largely solved this problem, by reducing the affect of misleading hairstyles, while not removing important hair information. Figure 5 shows examples of individual faces that are misclassified when the original full face images are used and classified correctly by the global and feature based model.

The global and feature based model for gender classification presented here performs significantly better than the global and featural models individually. This model allows inspection of facial data at various component levels and the results presented suggest that all components carry high levels of gender information. We believe that this type of representation also acts as a weighting factor of information, where highly variable discriminatory information (like hair) alone does not affect classification. Importantly, the global and feature based model captures an attentional component of human face recognition, whereby a human observer may use specific face feature cues to aid gender identification. Our experiments with human subjects showed that impressive levels of gender recognition accuracy were obtained from low resolution representations of single facial features (i.e., eyes and mouths). This underscores the importance of these specific features and supports the psychological plausibility of the global and feature based model discussed in this paper. Moreover, there was some association between the errors made by the models and those made by human observers. This, again, supports the psychological plausibility of these models although we will need to replicate this in some new sets of feature images that reflect a greater number of classification errors by the 3 models. We hope that this approach will also facilitate a useful comparison between the different dimensionality reduction techniques. Finally, we note that the Performance of CCA, a nonlinear technique, is comparable to PCA, with the added advantage that it uses far fewer variables than PCA.

Fig. 5. Examples of misclassified faces due to hair.

Acknowledgments

The authors are grateful to Kerry Foley for assistance in collecting data for human experiments.

References

1. R. E. Bellman, *Adaptive Control Processes: A Guided Tour* (Princeton University Press, 1961).
2. H. B. Barlow, Unsupervised learning, *Neural Computation* **1** (1989) 295–311.

3. I. T. Jolliffe, *Principal Component Analysis* (New York: Springer-Verlag, 1986).
4. P. Demartines and J. Herault, Curvilinear component analysis: A self-organizing neural network for nonlinear mapping of data sets, *IEEE Transactions on Neural Networks* **8** (1997) 148–154.
5. T. Kohonen, *Self Organizing Maps*, 3rd edn., (Springer-Verlag, 2001).
6. V. Bruce, A. M. Burton, E. Hanna, P. Healy, O. Mason, A. Coombes, R. Fright and A. Linney, Sex discrimination: How do we tell the difference between male and female faces?, *Perception* **22** (1993) 131–152.
7. A. M. Burton, V. Bruce and N. Dench, What's the difference between men and women? Evidence from facial measurement, *Perception* **22** (1993) 153–176.
8. H. Abdi, D. Valentin, B. Edelman and A. J. O'Toole, More about the difference between men and women: Evidence from linear neural networks and the principal component approach, *Perception* **24** (1995) 539–562.
9. B. A. Golomb, D. T. Lawrence and T. J. Sejnowski, Sexnet: A neural network identifies sex from human faces, *Advances in Neural Information Processing Systems* **3** (1991) 572–577.
10. R. Brunelli and T. Poggio, HyperBF networks for gender classification, *DARPA Image Understanding Workshop* (1992).
11. B. Moghaddam and M.-H. Yang, Gender classification with support vector machines, *Mitsubishi Electric Research Laboratory, Technical Report TR-2000-01* (2000).
12. Z. Sun, X. Yuan, G. Bebis and S. J. Louis, Neural-Network-based gender classification using genetic search for eigen-feature selection, *IEEE Int. Joint Conf. Neural Networks* (2002).
13. S. Buchala, N. Davey, R. J. Frank and T. M. Gale, Dimensionality reduction of face images for gender classification, *Department of Computer Science, The University of Hertfordshire, UK, Technical Report 408* (2004).
14. A. Luckman, N. M. Allinson, A. M. Ellis and B. M. Flude, Familiar face recognition: A comparative study of a connectionist model and human performance, *Neurocomputing* **7** (1995) 3–27.
15. P. J. Phillips, H. Wechsler, J. Huang and P. Rauss, The FERET database and evaluation procedure for face recognition algorithms, *Image and Vision Computing* **16** (1998) 295–306.
16. A. M. Martiniz and R. Benavente, The AR face database, *CVC, Technical Report 24* (1998).
17. O. Jesorsky, K. Kirchberg and R. Frischholz, Robust face detection using the hausdorff distance, *Int. Conf. Audio- and Video-based Biometric Person Authentification* (Halmstad, Sweden, 2001).
18. C. Cortes and V. Vapnik, Support-vector networks, *Machine Learning* **20** (1995) 273–297.
19. P. Grassberger and I. Proccacia, Measuring the strangeness of strange attractors, *Physica D* **9** (1983) 189–208.