GLOBAL AND FEATURE BASED GENDER CLASSIFICATION OF FACES: A COMPARISON OF HUMAN PERFORMANCE AND COMPUTATIONAL MODELS

SAMARASENA BUCHALA[A] TIM M.GALE[A,B] NEIL DAVEY[A] RAY J.FRANK[A]
KERRY FOLEY[B]

[A] *Department of Computer Science, University of Hertfordshire, College Lane, Hatfield, AL10 9AB, UK*
{S.Buchala, N.Davey, T.Gale, R.J.Frank}@herts.ac.uk

[B] *Department of Psychiatry, QEII Hospital, Welwyn Garden City, AL7 4HQ, UK*
kbmf@btopenworld.com

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.

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 (e.g. eyes, mouth etc). 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 present these two kinds of representation and use a Support Vector Machine (SVM) to classify gender. Since face image data has very high dimensionality, we also implement dimensionality reduction techniques before classification.

The remainder of the paper is organised as follows. Related work is discussed in the next Section. Section 3 discusses the methodology used for this study. Sections 4 and 5 present the computational and human experimental results. We conclude with some discussion of the results in Section 6.

2. Related Work

Issues in gender classification have stimulated a great deal of research by psychologists and computer scientists. While research in Psychology (Bruce et al., 1993, Burton et al., 1993, Abdi et al., 1995) has largely been within the context of human visual processing, and identifying key featural differences in males and females, Computer Science research (Golomb et al., 1991, Brunelli & Poggio, 1992, Moghaddam & Yang, 2000, Sun et al., 2002) 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.

3. Methodology

3.1. *Face Representation*

Hair, especially for females, forms a major part of a facial image and has a dominating effect on classification. Abdi et al (1995) reported gender classification accuracy of 80% for hairless faces against 91.8% for the same faces with hair information included. However, classification rate on hairless faces was better than that on faces with hair information in our previous work (Buchala et al., 2004). The performance degradation for face images with hair 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 Figure 1 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 Figure 1 are the pictorial views of the standard deviations within 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.

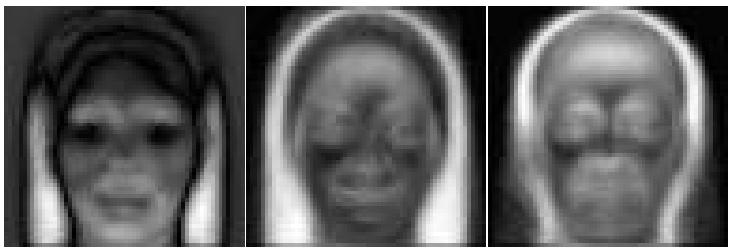

Figure 1. The first image is the pictorial representation of the difference of the means, of female and male face images. The second image is the standard deviation within the female face images. The third image is the standard deviation within the male face images.

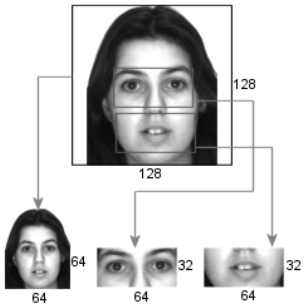

Figure 2. Three sub-images are obtained from the original 128 × 128 image. A 32 × 64 image pertaining to the eye region and a 32 × 64 image pertaining to the mouth region are extracted from the original image. The third sub-image is a 64 × 64 reduced resolution version of the original image.

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 Figure 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 (1995) for their computational model of familiar face recognition.

3.2. *Dimensionality Reduction*

Face image data has very high dimensionality and owing to the "curse of dimensionality" (Bellman, 1961), we apply dimensionality reduction techniques before using an SVM for classification.

Principal Component Analysis (PCA) (Jolliffe, 1986) is a popular dimensionality reduction technique that linearly transforms a D dimensional dataset X to a d dimensional dataset Y, without significant loss of information, where $d \leq D$.

Self Organising Map (SOM) (Kohonen, 2001) is a 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.

Curvilinear Component Analysis (CCA) (Demartines & Herault, 1997), a recent technique, has the ability to reduce the dimensionality of strongly-nonlinear data. The output is a free space that assumes the shape of the submanifold of the data. CCA minimizes the following error function:

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

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. $F_{\lambda}\left(d_{i,j}^{Y}\right)$ is the neighbourhood function. The idea of CCA is to match distances in the input and output spaces. However, preservation of larger distances may not be possible in the case of nonlinear data. In this case, it is important that at least local (smaller) distances should be preserved. For this reason CCA uses the neighbourhood function that ensures the condition of distance matching is satisfied for smaller distances while it is relaxed for larger distances.

3.3. *Support Vector Machine*

The classification is performed using an SVM. The SVM (Cortes & Vapnik, 1995) is a recently developed learning method, for pattern classification and regression. The basic idea of the SVM is to find the optimal hyperplane that has the maximal margin of separation between the classes, while having minimum classification errors.

Given a set of examples and their labels $\{(\mathbf{x_1}, \mathbf{y_1}), (\mathbf{x_2}, \mathbf{y_2}),\ldots(\mathbf{x_N}, \mathbf{y_N})\}$ where $\mathbf{y_i} \in \{-1,1\}$, the optimal hyperplane is given as:

$$f(X) = \sum_{i=1}^{N} \alpha_i y_i k(X, X_i) + b \tag{2}$$

Constructing the optimal hyperplane is equivalent to finding α_i with nonzero values. The examples corresponding to the nonzero α_i are called support vectors. $k(X, X_i)$ is a kernel function, which implicitly maps the example data points into a high dimensional feature space, and takes inner product in that feature space. The potential benefit of a kernel function is that the data is more likely to be linearly separable in the high dimensional feature space, and also the actual mapping to the higher-dimensional space is never needed.

4. 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 (Phillips et al., 1998), AR (Martiniz & Benavente, 1998), and BioId (Jesorsky et al., 2001). All face images are aligned based on their eye-locations. 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 reduce lighting effects. 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 classifier, with RBF kernel. Before applying classification, dimensionality reduction techniques discussed in Section 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[a] estimation technique, the Correlation Dimension (Grassberger & Proccacia, 1983), 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.

---

[a] Due to correlations, linear and nonlinear, a D dimensional data may actually lie in a d dimensional space. This true dimension d is called Intrinsic Dimension, where $d \leq D$. As PCA accounts only linear correlations, it is unable to reduce the data dimension to its intrinsic dimension when the correlations are nonlinear.

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 applying dimensionality reduction on the sub-images individually and combining the resultant data. It can be seen from Table 2 that PCA performed marginally better than CCA and 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 3 shows that the composite data outperformed all other data representations.

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% (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)

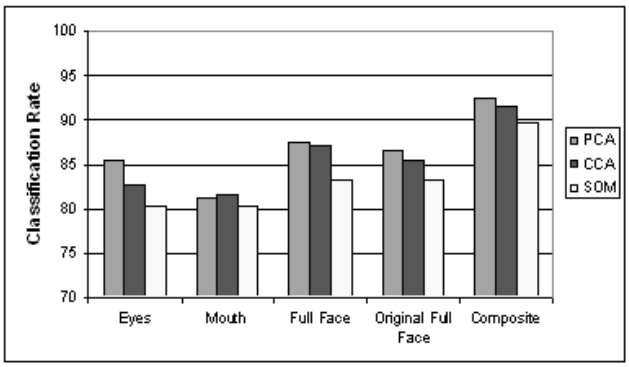

Figure 3. Average classification rates on different features.

5. Human Experiments

We recruited 80 participants (40 Male, 40 Female, mean age 40.1 years) to undertake a gender classification task using exactly the same sub-images that were used to test the dimensionality reduction techniques. Each participant viewed 80 eye and 80 mouth sub-images and were asked to record their best guess for the gender for each image. Each set (eyes and mouths) comprised 40 male and 40 female images, presented in random order. Data were collated by subjects and by items (the latter for the purpose of error analysis). We compared human and model performance on gender classification (N.B. in this context, the model performance was that for the sub images only, not the composite data).

5.1. *Eye Images*

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

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

Model	No. Incorrect Items	No. Correct Items	Mean human accuracy for items incorrectly classified by the model	Mean human accuracy for items correctly classified by the model
PCA	13	67	72.7 %	78.4%
CCA	13	67	71.9 %	78.6%
SOM	19	61	57.1 %	83.8%

Although the accuracy of humans was always higher for items that had been correctly classified by the models, this difference was statistically significant

only for the SOM ($p < 0.005$). Since the data were not normally distributed, differences were analysed non-parametrically (with Mann-Whitney's 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.

5.2. *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 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. These data are presented in table 4.

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

Model	No. Incorrect Items	No. Correct Items	Mean human accuracy for items incorrectly classified by the model	Mean human accuracy for items correctly classified by the model
PCA	15	65	57 %	79.7 %
CCA	20	60	54.6 %	82.4 %
SOM	21	59	59.2 %	81.2 %

The differences were significant at $p < 0.001$ or less for all 3 methods, showing that those items which the models fail to categorise correctly are more likely to elicit gender recognition errors in humans.

6. Discussion and Conclusion

Hair, especially for females, forms a major part of the image and has a dominating effect 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 effect of misleading hairstyles, while not removing important hair information. Figure 4 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.

Figure 4. Examples of the faces that are misclassified due to hair style of the individuals.

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.

References

Abdi, H., Valentin, D., Edelman, B., O'Toole, J. A. (1995). More about the difference between men and women: evidence from linear neural networks and the principal component approach. *Perception, 24*, 539-562.

Bellman, R. E. (1961). *Adaptive control processes: A guided tour*: Princeton University Press.

Bruce, V., Burton, A. M., Hanna, E., Healy, P., Mason, O., Coombes, A., Fright, R., Linney, A. (1993). Sex discrimination: how do we tell the difference between male and female faces? *Perception, 22*, 131-152.

Brunelli, R., Poggio, T. (1992). *HyperBF networks for gender classification.* Paper presented at the DARPA Image Understanding Workshop.

Buchala, S., Davey, N., Frank, R. J., Gale, T. M. (2004). *Dimensionality reduction of face images for gender classification* (Technical Report 408): Department of Computer Science, The University of Hertfordshire, UK.

Burton, A. M., Bruce, V., Dench, N. (1993). What's the difference between men and women? Evidence from facial measurement. *Perception, 22*, 153-176.

Cortes, C., Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*, 273-297.

Demartines, P., Herault, J. (1997). Curvilinear component analysis: A self-organizing neural network for nonlinear mapping of data sets. *IEEE Transactions on Neural Networks, 8*(1), 148-154.

Golomb, B. A., Lawrence, D. T., Sejnowski, T. J. (1991). Sexnet: A neural network identifies sex from human faces. *Advances in Neural Information Processing Systems, 3*, 572-577.

Grassberger, P., Proccacia, I. (1983). Measuring the strangeness of strange attractors. *Physica D, 9*, 189-208.

Jesorsky, O., Kirchberg, K., Frischholz, R. (2001). *Robust face detection using the hausdorff distance.* Paper presented at the International Conference on Audio- and Video-based Biometric Person Authentification, Halmstad, Sweden.

Jolliffe, I. T. (1986). *Principal component analysis.* New York: Springer-Verlag.

Kohonen, T. (2001). *Self organizing maps* (3rd ed.): Springer-Verlag.

Luckman, A., Allinson, N. M., Ellis, A. M., Flude, B. M. (1995). Familiar face recognition: A comparative study of a connectionist model and human performance. *Neurocomputing, 7*, 3-27.

Martiniz, A. M., Benavente, R. (1998). *The AR face database* (Technical Report 24): CVC.

Moghaddam, B., Yang, M.-H. (2000). *Gender classification with support vector machines* (Technical Report TR-2000-01): Mitsubishi Electric Research Laboratory.

Phillips, P. J., Wechsler, H., Huang, J., Rauss, P. (1998). The FERET database and evaluation procedure for face recognition algorithms. *Image and Vision Computing, 16*(5), 295-306.

Sun, Z., Yuan, X., Bebis, G., Louis, S. J. (2002). *Neural-Network-based gender classification using genetic search for eigen-feature selection.* Paper presented at the IEEE international joint conference on neural networks.