Gender Classification Using Principal Geodesic Analysis and Gaussian Mixture Models

Jing Wu, William A.P. Smith, and Edwin R. Hancock

Department of Computer Science, The University of York, York, YO10 5DD, UK
{jwu, wsmith, erh}@cs.york.ac.uk

Abstract. The aim in this paper is to show how to discriminate gender using a parameterized representation of fields of facial surface normals (needle-maps) which can be extracted from 2D intensity images using shape-from-shading (SFS). We makes use of principle geodesic analysis (PGA) to parameterize the facial needle-maps. Using feature selection, we determine which of the components of the resulting parameter vector are the most significant in distinguishing gender. Using the EM algorithm we distinguish gender by fitting a two component mixture model to the vectors of selected features. Results on real-world data reveal that the method gives gender discrimination results that are comparable to human observers.

1 Introduction

Humans are remarkably accurate determining the gender of a subject based on the appearance of the face alone. In fact, an accuracy as good as 96% can be achieved with the hair concealed, facial hair removed and no makeup [1]. Experiments by Bruce etc. showed that the gender of the face is conveyed by several cues including: (i) superficial and/or local features, (ii) configural relationships between features, and (iii) the 3-D structure of the face [2]. In [1], Burton etc. attempt to discover a gender discriminator by explicit measurements on the feature points of frontal facial views and profile views. However, their method requires manually labeled 14 landmark points. As a result it is unsuitable for automatic gender classification.

In this paper, we present a statistical framework for gender discrimination that does not require explicit landmark measurements. The method makes use of a representation of facial shape based on a parameterisation of fields of facial surface normals or needle-maps. The needle-map is a 2.5-D shape representation which is mid-way between the 2D intensity image and the 3D surface height function [3]. The representation can be acquired from 2D intensity images using shape-from-shading [4] and is invariant to illumination. To parameterise the facial needle-maps we make use of principle geodesic analysis (PGA) [5], [6]. PGA is a generalization of principle components analysis (PCA) [7]. For data residing on a Riemannian manifold, PGA is better suited to the analysis of directional data than PCA. Our aim is to determine gender using vectors of

J.F. Martínez-Trinidad et al. (Eds.): CIARP 2006, LNCS 4225, pp. 58–67, 2006.

PGA parameters. We aim to distinguish the genders of a sample of subjects by fitting a two-component mixture model to the distribution of vectors of selected features.

The standard method to learn the mixture models is the expectation - maximization (EM) algorithm [8], [9], [10], [11]. However, applying the EM algorithm directly to high dimensional facial needle-maps yields two problems. The first is the analysis of the distribution of needle-maps cannot be effected in a linear way, because a linear combination of unit vectors (normals) is not itself a unit vector. The second problem is that the covariance over the full dimensions of the data is too large to be computationally tractable.

The first of these problems is overcomed if we use PGA parameters (feature vectors) to represent the facial needle-maps since the parameter vectors reside in a vector space. To overcome the problem of dimensionality, we select the most significant feature components for discriminating gender that give the best class separability. Experimental results show that the mixture model learnt by our method has a correct gender classification rate of 87%.

The outline of the paper is as follows. Section 2 reviews the log and exponential maps used in principal geodesic analysis. Section 3 explores how the most significant gender features can be selected. In Section 4, a detailed description of the learning phase and the classification method is given. Experiments are presented in Section 5. Finally, Section 6 concludes the paper and offers directions for future investigation.

2 Principle Geodesic Analysis

The surface normal $n \in R^3$ may be considered as a point lying on a spherical manifold $n \in S^2$, therefore, we turn to the intrinsic mean and PGA proposed by Fletcher et al. [5] to analyze the variations of the surface normals.

2.1 The Log and Exponential Maps

If $u \in T_n S^2$ is a vector on the tangent plane to S^2 at n and $u \neq 0$, the exponential map, denoted Exp_n, of u is the point, denoted $Exp_n(u)$, on S^2 along the geodesic in the direction of u at distance $\parallel u \parallel$ from n. This is illustrated in Fig. 1. The log map, denoted Log_n is the inverse of the exponential map. The exponential and log maps reserve the geodesic distance between two points, i.e. $d(n_1, n_2) = d(u_1, u_2)$, where $u_1 = Log_n n_1, u_2 = Log_n n_2$.

2.2 Spherical Medians

It is more natural to treat the surface normal as points on a unit sphere: $n_1, \ldots n_N \in S^2$ rather than points in Euclidian space. Instead of the Euclidian mean, we compute the intrinsic mean: $\mu = \arg\min_{n \in S^2} \sum_{i=1}^{N} d(n, n_i)$, where $d(n, n_i) = \arccos(n \cdot n_i)$ is the arc length. For a spherical manifold,

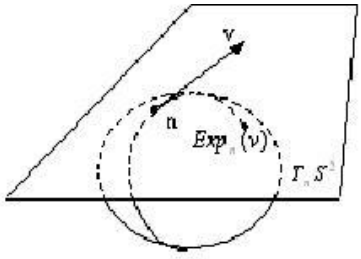

Fig. 1. The exponential map

the intrinsic mean can be found using the gradient descent method of Pennec [6]. Accordingly, the current estimate $\mu^{(t)}$ is updated as follows: $\mu^{(t+1)} = \mathrm{Exp}_{\mu^{(t)}}(\frac{1}{N}\sum_{i=1}^{N}\mathrm{Log}_{\mu^{(t)}}(n_i))$.

2.3 PGA of Needle Maps

PGA is analogous to PCA except that each principal axis in PCA is a straight line, while in PGA each principle axis is a geodesic curve. In the spherical case this corresponds to a great circle. Consider a great circle G on the sphere S^2. To project a point $n_1 \in S^2$ onto a point on G, we use the projection operator $\pi_G : S^2 \longrightarrow G$ given by $\pi_G(n_1) = \mathrm{argmin}_{n\in G}(n_1, n)^2$. For a geodesic G passing through the intrinsic mean μ, π_G may be approximated linearly in the tangent plane $T_\mu S^2$: $\mathrm{Log}_\mu(\pi_G(n_1)) \approx \sum_{i=1}^{K} V^i \cdot \mathrm{Log}_\mu(n_1)$, where $V_1, \ldots V_K$ is an orthonormal basis for $T_\mu S^2$.

Suppose there are K training needle-maps each having N pixel locations, and the surface normal at the pixel location p for the k^{th} training needle-map is n_p^k. We calculate the intrinsic mean μ_p of the distribution of surface normals $n_p^1, \ldots n_p^K$ at each pixel location p. n_p^k is then represented by its log map position $u_p^k = \mathrm{Log}_{\mu_p}(n_p^k)$. $u^k = [u_1^k, \ldots, u_N^k]^T$ is the log mapped long vector of the k^{th} training needle-map. The K long vectors form the data matrix $U = [u^1|\ldots|u^K]$. The covariance matrix of the data matrix is $L = \frac{1}{K}UU^T$.

We use the numerically efficient snap-shot method of Sirovich [12] to compute the eigenvectors of L. Accordingly, we construct the matrix $\hat{L} = \frac{1}{K}U^TU$, and find the eigenvalues and eigenvectors. The i^{th} eigenvector e_i of L can be computed from the i^{th} eigenvector $\hat{e_i}$of $\hat{L}$ using $e_i = U\hat{e_i}$. The i^{th} eigenvalue λ_i of L equals the i^{th} eigenvalue $\hat{\lambda_i}$ of $\hat{L}$ when $i \leq K$. When $i > K$, $\lambda_i = 0$. Providing the effects of noise are small, we only need to retain S eigenmodes to retain p percent of the model variance. S is the smallest integer satisfying:$\sum_{i=1}^{S}\lambda_i \geq \frac{p}{100}\sum_{i=1}^{K}\lambda_i$. In our experiments, we use the 10 leading eigenvectors of L as the columns of the eigenvector matrix (projection matrix) $\Phi = (e_1|e_2|\ldots|e_{10})$.

Given a long vector $u = [u_1, \ldots, u_N]^T$, we can get the corresponding vector of parameters (feature vector) $b = \Phi^T u$. Given a feature vector $b = [b_1, \ldots b_S]^T$, we can generate a needle-map using: $n_p = \mathrm{Exp}_{\mu_p}((Pb)_p)$.

3 Feature Selection

After PGA, we select the most significant S components of the PGA parameter vector (in our experiments, S=10). However, the S dimensional feature vectors still inevitably contain information which is either redundant or irrelevant to the gender classification task. As stated in [13], the classification of patterns as performed by humans is based on a very few of the most important attributes. Therefore, we select the most significant features for gender discrimination.

We examine the distribution for the first 9 components of the PGA parameter vectors for the 200 data samples in our experiments. Here the first 100 are females, and the last 100 are males (see Fig. 2). Table 1 shows the mean values of the first 9 feature components for females and males. By inspection, the 1st, 5th and 6th components have the most significant difference between females and males.

Figure 3 shows the mean face and its variations along the directions of the 1st, 5th and 6th principal geodesic directions. We can see the 3 feature components do convey some gender information. Turning our attention to the 1st component, as λ_1 increases, the face becomes larger and more solid, and, the cheeks thinner. These are all masculine characteristics. In the case of the 5th component, as λ_5 decreases, the face becomes wider and the eyes deepen. Again these are masculine characteristics. In the case of the 6th component, as λ_6 increases, there is a more masculine appearance. Fig. 2 and Fig. 3 therefore indicate the 1st, 5th, 6th features are intuitively the most significant ones for gender discrimination.

To verify our empirical selection, we explore the different feature selection criteria described by Devijver and Kittler [13]. We use the class separability criterion $J(\xi) = \frac{|S_w+S_b|}{|S_w|} = \prod_{k=1}^{d}(1+\lambda_k)$, where S_w and S_b are the between and within class scatter matrices, λ_k, $k = 1 \ldots d$ are the eigenvalues of matrix $S_w^{-1}S_b$. The values of J for the first 9 features are shown in Table 2, from which we can see B(1), B(5), B(6) have the 3 largest values. The result is consistent with our empirical selection. Therefore, the 1st, 5th, and 6th features are the most significant features for gender discrimination. We use them as the selected feature vectors in gender classification using EM algorithm.

Table 1. Mean values of the first 9 feature components

	B(1)	B(2)	B(3)	B(4)	B(5)	B(6)	B(7)	B(8)	B(9)
Female	-8.6776	-3.2660	-0.4854	0.7371	3.0581	-2.4635	0.1951	0.2868	0.8078
Male	8.6776	3.2660	0.4854	-0.7371	-3.0581	2.4635	-0.1951	-0.2868	-0.8078

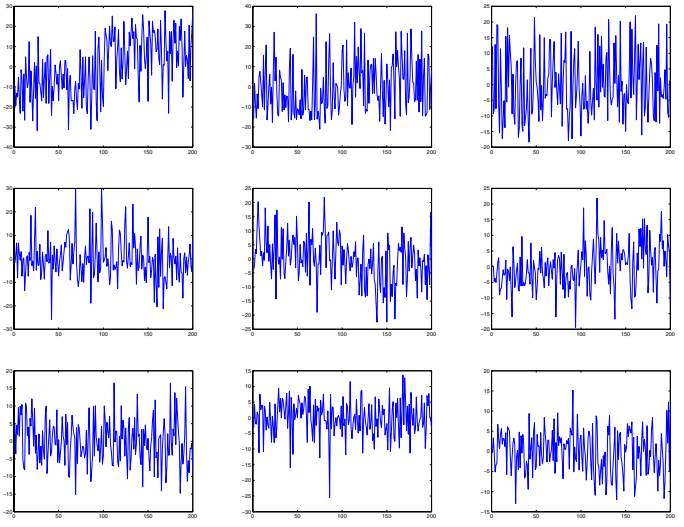

Fig. 2. Plots of the first 9 feature components. From left to right, the first line is B(1), B(2), B(3), the second line is B(4), B(5), B(6), the third line is B(7), B(8), B(9). X axis ranging from 1 to 200 stand for the 200 faces, the first 100 are females, the second 100 are males. Y axis ranging from -25 to 25 is the value of the feature components.

Fig. 3. The mean face and its variances along the 1st, 5th and 6th feature components. The lines are according to the features (from top to bottom): 1st, 5th, 6th features. The columns are according to the variances (from left to right): λ=-30, λ=-20, λ=0 (the mean face), λ=20, and λ=30.

Table 2. J values on the first 9 feature components

	B(1)	B(2)	B(3)	B(4)	B(5)	B(6)	B(7)	B(8)	B(9)
J	1.6231	1.0681	1.0024	1.0073	1.1644	1.1532	1.0010	1.0026	1.0243

4 Learning Gaussian Mixture Models

We use the EM algorithm to fit a two component mixture model to vectors of selected features, and explore whether the a posteriori class probabilities can be used to classify the gender of subjects.

4.1 EM Initialization

In our EM algorithm the a posteriori probability is estimated from the selected feature vectors using the method outlined in [10]. For 2-component Gaussian mixture models, we set $\alpha_1^{(0)} = \alpha_2^{(0)} = \frac{1}{2}$, $\mu_{b1}^{(0)} = \mu_b + \varepsilon_1$, $\mu_{b2}^{(0)} = \mu_b + \varepsilon_2$, and $\Sigma_{b1}^{(0)} = \Sigma_{b2}^{(0)} = det(\Sigma_b)^{1/d} I_d$. Here α_1, α_2 is the a priori probability of each class, μ_b is the overall mean of the selected feature vectors, Σ_b is the overall covariance matrix of the selected feature vectors, ε_1 and ε_2 are two small random vectors.

In our experiments $d = 3$. We can set the class means are $\mu_{b1}^{(0)} = [-\varepsilon_1(1), +\varepsilon_1(2), -\varepsilon_1(3)]$, $\mu_{b2}^{(0)} = [+\varepsilon_2(1), -\varepsilon_2(2), +\varepsilon_1(3)]$. The signs before the ε elements are indicated from Fig. 3. We can see, in our experiments, when the 1st and 6th feature components are negative, the 5th component is positive, the face is more female. Otherwise, the face is more male. Using this information makes the EM initialization more reliable.

4.2 E – Step

In E – Step the a posteriori class membership probability is updated by applying the Bayes law to the class-conditional density. In our application, the class-conditional density is Gaussian:
$p(B_j|\mu_{bc}^{(t)}, \Sigma_{bc}^{(t)}) = \frac{1}{\sqrt{(2\pi)^d |\Sigma_{bc}^{(t)}|}} \exp[-\frac{1}{2}(B_j - \mu_{bc}^{(t)})^T \times (\Sigma_{bc}^{(t)})^{-1} \times (B_j - \mu_{bc}^{(t)})]$. Here, B_j donates the selected feature vector of the jth sample data.

At iteration t+1, the a posteriori probability is updated as follows:

$$W_c^{(j,t)} \equiv P(j \in c|B_j, \mu_{bc}^{(t)}, \Sigma_{bc}^{(t)}) = \frac{\alpha_c^{(t)} p(B_j|\mu_{bc}^{(t)}, \Sigma_{bc}^{(t)})}{\sum_{i=1}^{2} \alpha_i^{(t)} p(B_j|\mu_{bi}^{(t)}, \Sigma_{bi}^{(t)})}. \quad (1)$$

Here, $W_c^{(j,t)}$ means estimate, at iteration t, of the probability that B_j was produced by class c.

4.3 M – Step

In M – Step the parameters for each class are updated to maximize the expected log-likelihood function:
$Q(C^{(t+1)}|C^{(t)}) = \sum_{j=1}^{n} \sum_{c=1}^{2} W_c^{(j,t)} \times \log(\alpha_c^{(t+1)} p(B_j|\mu_{bc}^{(t+1)}, \Sigma_{bc}^{(t+1)}))$.

At iteration t+1, the revised estimate of the a priori probability of class c is $\alpha_c^{(t+1)} = \frac{1}{n}\sum_{j=1}^{n} W_c^{(j,t)}$, the revised estimate of the mean vector is $\mu_{bc}^{(t+1)} = \frac{\sum_{j=1}^{n} W_c^{(j,t)} B_j}{\sum_{j=1}^{n} W_c^{(j,t)}}$, and the revised estimate of the covariance matrix is $\Sigma_{bc}^{(t+1)} = \frac{\sum_{j=1}^{n} W_c^{(j,t)} (B_j - \mu_{bc}^{(t+1)})(B_j - \mu_{bc}^{(t+1)})^T}{\sum_{j=1}^{n} W_c^{(j,t)}}$.

4.4 Classification

After the mixture model of genders has been learnt, we use the a posteriori class probability to classify faces to one of the genders. Given the needle-map n of a test face, first get its selected feature vector b through PGA and feature selection method mentioned in previous sections. Then compute the a posteriori probabilities W_f and W_m through formula (1), using the acquired mean vectors μ_{bf}, μ_{bm} and the covariance matrixes Σ_{bf}, Σ_{bm}. If $W_f > W_m$, then the face is classified as female. Otherwise, the face is classified as male.

5 Experiments

In this section, we evaluate the performance of the method for discriminating gender on the basis of the learnt two-component mixture model for the distribution of shape-features. The data consists of 200 facial needle-maps with known ground truth from the Max Plank dataset. There are 100 females and 100 males.

We first apply PGA and feature selection to the data to extract the shape parameter vectors and perform feature selection. The visualization of the data is shown in the left-hand panel of Fig. 5. Here we show the distribution of the 1st and 5th features as a scatter plot. The data is relatively well clustered according to gender. There is some overlap and this is due to feminine looking males and masculine looking females.

Experiment 1. We use the 200 data for unsupervised learning. Figure 4 shows the initial and final classifications of the data. After convergence of the EM algorithm, the data are reasonably well clustered according to gender. The correct classification rate reaches 89% for females, and 85% for males. Figure 5 visualizes the classification results. From the figure, around 13% of the errors are due to the misclassification of the data in the overlap region. Tests involving human observers give similar error rates.

Experiment 2. We select the 10 needle-maps with the largest and 10 with the smallest female probability W_f from the 100 female faces. The top 10 faces are considered to be typical females, while the bottom 10 are considered to be female faces falling into the overlap region. We repeat this procedure for the male faces. We render the 40 facial needle-maps with facial textures, and present them to 9 subjects (6 males and 3 females). The average classification error rate of the 9 people is shown in Table 3. From the table, the classification performance on

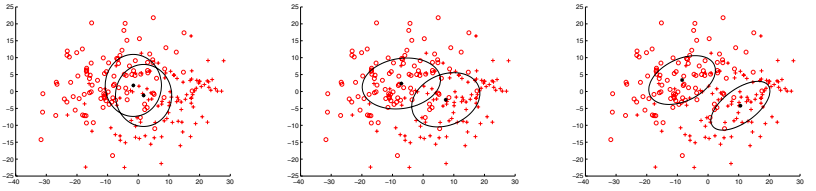

Fig. 4. Learning steps visualized on 1st and 5th features. From left to right, are the results of EM initialization, after 5 iterations and on convergence.

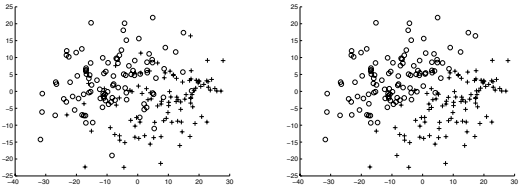

Fig. 5. Classification result on the training data visualized on 1st and 5th features. Left is the original training data, right is the classification result.

the faces in the overlap region is poorer than that of the typical female and male faces. This confirms that the results obtained in experiment 1 are consistent with the performance of human subjects. Interestingly, the classification of the female faces is poorer than that of the males. This may be due to the fact that without hair or makeup the facial appearance is masculine.

Table 3. Error rate of human classification

Total	Overlap	Unoverlap	Females	Males
22.5%	25.6%	19.4%	43.9%	1.1%

Experiment 3. We randomly selected 40 needle-maps from the 200 available for use as test data. The remaining 160 are used as training data. First, we obtain the selected feature vectors of the training and test data using the intrinsic mean and projection matrix using the full sample of 200 data. Then we fit the mixture model to the training data. We visualize the models in the left-hand panel of Fig. 6. The classification rate is evaluated by fitting the mixture models to the test data. The result is shown in the right-hand panel of Fig. 6 and compared with the original test data shown in the middle of Fig. 6. The classification rate for females is 80% and on males 95%. This is a good result and that our method has good generalisation.

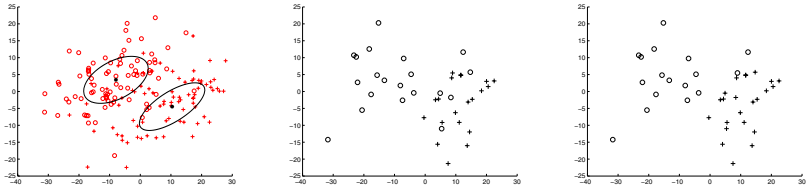

Fig. 6. Training and Testing result visualized on 1st and 5th features. The left is the learnt models on the training data. In the middle is the original testing data. The right is the classification on the testing data.

6 Conclusion

In this paper, we apply feature selection and EM algorithm to PGA shape parameters of the facial needle-maps to perform gender classification. We explore the most significant components of the parameter vectors for gender discrimination, and use the EM algorithm to cluster the selected features. Experimental results show that using the selected feature vectors, facial needle-maps can be well clustered according to gender. Moreover, it demonstrates that it is feasible to construct a 2-component Gaussian mixture models from the facial needle-maps to classify the gender.

However, there are still some problems that require further investigation. First, feature selection is quite empirical and the simplest theoretical verification measure has been used. Our future research will focus on the use of more principled methods for feature selection. Second, in the EM initialization step, we need to analyze the mean face and its variances along each feature component to determine the signs of the initial mean vectors. Thus, although the training data need not to be labeled, the learning phase is not totally unsupervised. Third, our current experiments are based on the ground truth needle-maps extracted from range images. In the future, we will apply our method to needle-maps recovered from facial images using SFS.

References

1. A Mike Burton, Vicki Bruce, Neal Dench: What's the difference between men and women? Evidence from facial measurement. Perception, vol.22, pp.153-176, 1993
2. Vicki Bruce, A Mike Burton, Elias Hanna, Pat Healey and Oli Mason, Anne Coombes, Rick Fright, Alf Linney: Sex discrimination: how do we tell the difference between male and female faces?. Perception, vol.22, pp.131-152, 1993
3. D.Marr: Vision. San Francisco: W.H. Freeman, 1982
4. William Smith , Edwin R. Hancock: Recovering Facial Shape and Albedo using a Statistical Model of Surface Normal Direction. Tenth IEEE International Conference on Computer Vision, vol.1, pp.588-595, 2005
5. P.T. Fletcher,S. Joshi,C. Lu,S.M. Pizer: Principal geodesic analysis for the study of nonlinear statistics of shape. IEEE Transactions on Medical Imaging, vol.23, pp.995-1005, 2004

6. X. Pennec: Probabilities and statistics on riemannian manifolds: A geometric approach. Technical Report RR-5093, INRIA, 2004
7. I.T. Jolliffe: Principle Component Analysis. Springer-Verlag, New York, 1986
8. M.A.T. Figueiredo,A.K. Jain: Unsupervised Learning of Finite Mixture Models. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.24,No.3,March 2002
9. M.A.T. Figueiredo,J.M.N. Leitao,A.K. Jain: On Fitting Mixture Models. Engergy Minimization Methods in Computer Vision and Pattern Recognition,E. Hancok and M. Pellilo, Eds: Springer-Verlag, pp. 54-69, 1999
10. Naonori Ueda,Ryohei Nakano,Zoubin Ghabramani,Geoffrey E.Hinton: SMEM Algorithm for Mixture Models. Neural Computation, vol.12(9), pp.2109-2128, 2000
11. Christophe Biernacki,Gilles Celeux,Gerard Govaert: Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. Computational Statistics & Data Analysis, vol.41, Issue 3-4, pp. 561-575, 2003
12. L. Sirovich: Turbulence and the dynamics of coherent structures. Quart. Applied Mathematics, vol.XLV, no. 3,pp.561-590, 1987
13. P. Devijver,J. Kittler: Pattern Recognition: A Statistical Approach. PrenticeHall, 1982