

An Information Theoretic Approach to Gender Feature Selection

Zhihong Zhang, Edwin R.Hancock*
University of York
York, UK
{zhihong, erh}@cs.york.ac.uk

Jing Wu
University of Cardiff
Cardiff, UK
J.Wu@cs.cardiff.ac.uk

Abstract

Most existing feature selection methods focus on ranking features based on an information criterion to select the best K features. However, several authors have found that the optimal feature combinations do not give the best classification performance [8],[7]. The reason for this is that although an individual feature may have limited relevance to a particular class, when taken in combination with other features it can be strongly relevant to the class. To overcome this problem, we draw on recent work on the graph embedding formulation of subspace learning where the projection matrix is constrained to be selection matrix [14] designed to select the optimal feature subset. In this paper, we derive a trace ratio (TR) criterion which selects features using a subset-level score rather than a feature-level score to perform feature selection. We apply the method to the challenging problem of gender determination using features delivered by principal geodesic analysis (PGA). A variational EM (VBEM) algorithm is used to learn a Gaussian mixture model on the selected feature subset and this is used to design a classifier for gender determination. We obtain a classification accuracy as high as 95% on 2.5D facial needle-maps, demonstrating the effectiveness of our feature selection method.

1. Introduction

High-dimensional data pose a significant challenge for gender classification. The most popular methods for reducing dimensionality are variance based subspace methods such as principal component analysis (PCA) [17]. PCA is an eigenvector method designed to model linear variation in high-dimensional data. Buchala et al. [4] have applied PCA to face images and explored which PCA components were the most discriminating, using linear discriminant analysis (LDA). Graf and Wichmann [10] applied PCA and LLE to

3D range images of human heads, and used SVM as the classifier for gender classification. Recently, a number of authors have found that the face images possibly reside on a nonlinear sub-manifold [5] [13]. However, the PCA fails to discover the underlying structure, if the face images lie on a nonlinear sub-manifold hidden in the image space. Some nonlinear techniques have been proposed to discover the nonlinear structure of the manifold, e.g., Wu et al. [18] explored the use of principal geodesic analysis (PGA) to recover fields of facial surface normals (needle maps) from brightness images using a model-based shape-from-shading technique [16]. Principal geodesic analysis (PGA) is an extension of PCA from Euclidean data to non-Euclidean data which reside on a non-linear manifold and is hence applicable to the analysis of directional surface normal data.

However, we found that the extracted PGA feature vectors only capture sets of features with a significant combined variance, it takes into account few considerations about the special characters and manifold structures of face images, and this renders them relatively ineffective for classification tasks. As a result, PGA is only suitable for dimension reduction and the PGA feature vectors probably contain information that is redundant or irrelevant to gender determination, and this limits the gender classification accuracy. Hence it is crucial to identify a smaller subset of features that are informative for classification and clustering. The idea underpinning feature selection is to select the features that are most relevant to classification while reducing redundancy. Recently mutual information (MI) has been shown to provide a principled way of measuring the mutual dependence of two variables, and has been used by a number of researchers to develop information theoretic feature selection criteria. For example, Batti [2] has developed the Mutual Information-Based Feature Selection (MIFS) criterion, where the features are selected in a greedy manner. Given a set of existing selected features S , at each step it locates the feature x_i that maximize the relevance to the class $I(x_i; C)$. The selection is regulated by a proportional term $\beta I(x_i; S)$ that measures the overlap information between the candidate feature and the existing features. The

*Edwin Hancock is supported by the EU FET project SIMBAD and by a Royal Society Wolfson Research Merit Award.

parameter β may significantly affect the features selected, and its control remains an open problem. Peng et al [15] on the other hand, use the so-called Maximum-Relevance Minimum-Redundancy criterion (MRMR), which is equivalent to MIFS with $\beta = \frac{1}{n-1}$. Yang and Moody’s [19] Joint Mutual Information (JMI) criterion is based on conditional MI and selects features by checking whether they bring additional information to an existing feature set. This method effectively rejects redundant features. Kwak and Choi [12] improve MIFS by developing MIFS-U under the assumption of a uniform distribution of information for input features. The method calculates the MI based on a Parzen window, which is less computationally demanding and also provides better estimates.

However, there are two limitations for the above MI feature selection methods. Firstly, they assume that every individual relevant feature should be dependent with a target class. This means that if a single feature is considered to be relevant it should be correlated with the target class, otherwise the feature is irrelevant [6]. So only a small set of relevant features is selected, and larger feature combinations are not considered. Hence, although a single feature may not be relevant, when combined with additional features it can become strongly relevant. The second weakness is that most of the existing methods simply consider pairwise feature dependencies, and do not check for third or higher order dependencies between the candidate features and the existing features. For example, in [1] there are four features X_1, X_2, X_3, X_4 , the existing selected feature subset is $\{X_1, X_4\}$, assume $I(X_2, C) = I(X_3, C)$, $I(X_2, X_1|C) = I(X_3, X_1|C)$, $I(X_2, X_4|C) = I(X_3, X_4|C)$ and $I(X_1, X_4, X_2) \gg I(X_1, X_2) + I(X_4, X_2)$, which indicates that X_2 has strong affinity with the joint subset $\{X_1, X_4\}$, although it has smaller individual affinity to each of them. So in this situation, X_2 may be discarded, and X_3 is selected, although the combination $\{X_1, X_4, X_2\}$ can produce a better cluster than $\{X_1, X_4, X_3\}$.

To overcome the above problem, in this paper, we adopt a graph-embedding view of feature selection by subspace learning. The method selects the feature subset based on a subset-level score rather than a feature-level score. It consists of three steps, namely, i) we first use PGA to find the principal geodesics that capture the largest statistical variance in the 2.5D facial needle-maps. Then each facial needle-map is represented by a parameter vector, referred to as PGA feature vector, ii) we then construct two weighted undirected graphs G_w and G_b in which each node corresponds to each facial needle-maps which is represented by the PGA feature vectors, iii) we further refine the extracted PGA feature vectors by employing a trace ratio (TR) criterion over the graph to locate the optimal feature set.

2. PGA on 2.5D Facial Needle-maps

Principal geodesic analysis (PGA) [9], is a generalization of principal component analysis (PCA) from the Euclidean space to Riemmanian manifolds. The goal of PCA is to find a linear subspace of the space in which the data lies, and maximize the variance of the projected data in the subspace. In PGA, the notion of a linear subspace is replaced by that of a geodesic sub-manifold. A surface normal n can be represented as a point residing on a spherical manifold $n \in S^2$. A facial needle-map is a field of N facial surface normals. It therefore may be considered as a point on the manifold $S^2(N) = \prod_{i=1}^N S^2$, which is a special case of a non-linear Riemmanian manifold. In Fig. 1, given $u \in T_n S^2$, a non-zero vector on the tangent plane to S^2 at the point $n \in S^2$, there exists a unique geodesic passing through n in the direction of u . The exponential map, denoted Exp_n , maps u to the point, denoted $Exp_n(u)$, on the geodesic in the direction u at distance $\|u\|$ from n . The log map, denoted Log_n is the inverse of the exponential map. Making use of exponential/log maps, the geodesic distance between two points n_1 and n_2 can be expressed as $dist_G(n_1, n_2) = \|Log_{n_1}(n_2)\|$, where $dist_G(\cdot, \cdot)$ is the geodesic distance between two points. PGA makes use of the intrinsic means of the data on the manifold and Lie group representation based on log and exponential maps. The intrinsic mean is defined as $\mu = argmin_{n \in S^2} \sum_{i=1}^N dist_G(n, n_i)$. To apply the PGA to facial needle-maps, we first make use of the log map to obtain the long vector representation of the faces in the tangent plane passing through the intrinsic means. Then, we compute the eigenvectors and the according eigenvalues of the covariance of the long vectors. The leading d eigenvectors from the projection matrix $\Phi = (e_1|e_2|\dots|e_d)$. Given a facial needle-map, we first obtain its long vector representation $u = [u_1, \dots, u_N]$ in the tangent plane, and then we represent the face using its PGA feature vectors $b = \Phi^T u$.

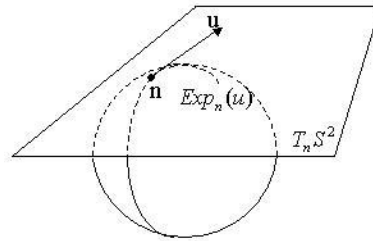


Figure 1. The exponential map

3. Learning Gender Discriminating Features

After principal geodesic analysis, each face is represented by its d dimensional PGA feature vector. However, not all of the PCA components are relevant to gender classification [4]. The irrelevant or redundant information limits

classification accuracy. As a result it is advantageous to select the most effective gender discriminating feature components from the PGA feature vectors. We make use of trace ratio (TR) criterion which selecting the features based on a subset-level score rather than a feature-level score.

3.1. Graph-Embedding View of Feature Selection

The idea underpinning feature selection is to a) reduce the dimensionality of the feature space, b) speed up and reduce the cost of a learning algorithm, c) obtain the feature subset which is most relevant to classification. In this paper, we demonstrate that these three tasks can be solved using graph embedding method. The purpose of graph embedding is to represent each vertex of the graph as a low dimensional vector that preserves similarities between the vertex pairs. The selected feature vector \bar{y} is given by $\bar{y} = W^T x$ where W^T is a column-full-rank matrix (usually determined by the graph Laplacian and adjacent matrix). The embedding can also be viewed as feature extraction. Feature selection differs from feature extraction, since it attempts to select the optimal feature subset in the original space. Graph embedding can be used to discover both the geometrical and the discriminative structure of the data manifold, where the projection matrix W^T play the role of a selection matrix Ψ^T .

$$y = \Psi^T x. \quad (1)$$

where $\Psi \in R^d$ is a selection matrix. Let Ψ_i denote a m -dimensional column vector:

$$\Psi_i = [0, \dots, 0, 1, 0, \dots, 0]. \quad (2)$$

The binary selection matrix Ψ is defined as

$$\Psi = [\Psi_{I(1)}, \Psi_{I(2)}, \dots, \Psi_{I(m)}]. \quad (3)$$

where the vector I is a permutation of $\{1, 2, \dots, d\}$ and d is the dimension of PGA feature vector.

Suppose there are n facial needle-maps. The data available to us is the set of PGA feature-vectors for the needle maps, $X = \{x_1, x_2, \dots, x_n\}$, where $x_i = (F_1^i, F_2^i, \dots, F_d^i)$ is the d dimensional feature-vector for the i -th needle-map. In order to discover both geometrical and discriminant structure of the data manifold, we construct two weighted graphs to capture the similarity structure of the data. The first is the intra-class or within class similarity graph $G_w(X, W_w)$, while the second is the inter or between class similarity graph $G_b(X, W_b)$. We employ the trace ratio (TR) criterion [14] over the graphs to locate the optimal feature subset.

The within-class similarity graph G_w is characterized by the weight matrix W_w and reflects the interclass compactness of the data, while G_b can be regarded as a between class penalty graph, characterized by the weight matrix W_b

which reflects the intraclass separability. Each facial needle map (represented by its PGA feature vector) is denoted by a vertex in the weighted graph. The two weight matrices $(W_w)_{ij}$ and $(W_b)_{ij}$ are respectively determined by the within class and between class pairwise similarity of the PGA vectors. When $(W_w)_{ij}$ is large, this implies that data x_i and data x_j belong to same class and a small value indicates they belong to different classes. Similarly, since $(W_b)_{ij}$ represents the global between class affinity relationships in the data, it provides a heavy penalty if data x_i and x_j belong to different classes. These features can be captured if the weight matrices W_w and W_b are defined as follows

$$(W_w)_{ij} = \begin{cases} \frac{1}{n_{c(i)}}, & \text{if } c(i) = c(j); \\ 0, & \text{if } c(i) \neq c(j). \end{cases} \quad (4)$$

$$(W_b)_{ij} = \begin{cases} \frac{1}{n} - \frac{1}{n_{c(i)}}, & \text{if } c(i) = c(j); \\ \frac{1}{n}, & \text{if } c(i) \neq c(j). \end{cases} \quad (5)$$

where $c(i)$ represents class label of data point x_i , and n_i denotes the number of data in class i .

We work with the Laplacian matrices for the graphs G_w and G_b . To this end let D_b and D_w denote the diagonal matrices of W_b and W_w , where $(D_b)_{ii} = \sum_{k=1}^n (W_b)_{ik}$ and $(D_w)_{ii} = \sum_{k=1}^n (W_w)_{ik}$. The weighted within-class degree of node i , i.e. $(D_w)_{ii}$ provides a natural measure of the density of data in the proximity of the data point x_i . Since the more data points that are close to x_i , the larger the weighted degree $(D_w)_{ii}$, the more important the point x_i . From the weight matrices W_b and W_w , and the degree matrices D_b and D_w , the corresponding between class and within class Laplacian matrices are $L_b = D_b - W_b$ and $L_w = D_w - W_w$ respectively. The optimal feature subsets should be the those for which the within class affinities are large, while the between class separation is large. These features are captured by selecting the set of features that minimize $\sum_{ij} \|x_i - x_j\|^2 (W_w)_{ij}$ and while maximizing $\sum_{ij} \|x_i - x_j\|^2 (W_b)_{ij}$.

3.2. Trace Ratio Criterion of Feature Selection on Subset-level

Given the n facial needle-maps $X = \{x_1, x_2, \dots, x_n\}$, each data is a high-dimensional data, which is represented by their d dimensions of PGA feature vectors $\{F_1, F_2, \dots, F_d\}$. To achieve the above tow objective functions, an appropriate criterion could be defined by:

$$\begin{aligned} \max \frac{\sum_{i \neq j} \|\Psi^T(x_i - x_j)\|^2 W_{b,ij}}{\sum_{i \neq j} \|\Psi^T(x_i - x_j)\|^2 W_{w,ij}} = \\ \max \frac{\text{tr}(\Psi^T X L_b X^T \Psi)}{\text{tr}(\Psi^T X L_w X^T \Psi)}. \end{aligned} \quad (6)$$

For the sake of simplicity, we denote $B = XL_bX^T$ and $E = XL_wX^T$. Suppose the subset-level score in Equation(6) reaches the global maximum λ^* if $\Psi_I = \Psi_{I^*}$, that is to say,

$$\max \frac{\text{tr}(\Psi_{I^*}^T B \Psi_{I^*})}{\text{tr}(\Psi_{I^*}^T E \Psi_{I^*})} = \lambda^* . \quad (7)$$

and

$$\max \frac{\text{tr}(\Psi_I^T B \Psi_I)}{\text{tr}(\Psi_I^T E \Psi_I)} \leq \lambda^*, \forall I . \quad (8)$$

From Equation(8), we can derive that

$$\max \text{tr}(\Psi_I^T (B - \lambda^* E \Psi_I)) \leq 0, \forall I . \quad (9)$$

Note that $\text{tr}(\Psi_{I^*}^T (B - \lambda^* E \Psi_{I^*})) = 0$ and let

$$f(\lambda) = \max \text{tr}(\Psi_I^T (B - \lambda E \Psi_I)), \forall I . \quad (10)$$

then we have $f(\lambda^*) = 0$. As $f(\lambda)$ is a monotonically decreasing function, finding the global optimal λ^* can be converted into the problem of locating the single root of equation $f(\lambda) = 0$. Here, we define score of the i -th feature as

$$\text{score}(F_i) = \Psi_i^T (B - \lambda E) \Psi_i . \quad (11)$$

The function $f(\lambda)$ can be rewritten as

$$f(\lambda) = \max \sum_{i=1}^m \Psi_{I(i)}^T (B - \lambda E) \Psi_{I(i)}, \forall I . \quad (12)$$

Thus $f(\lambda)$ equals to the sum of the first m largest scores.

The task of subset-level based feature selection is to seek the feature subset with the maximum score according to Equation(7). The root can be located using an iterative procedure to update λ and thus find the root of equation $f(\lambda) = 0$.

4. Classification

After finding the gender discriminating PGA feature components, we apply the variational EM (VBEM) algorithm [3] to fit a mixture of Gaussians model to the selected feature subset. After learning the mixture model, we use the a posteriori probability, see Equation(13), to classify faces with respect to gender. Given the needle-map of a test face, we first compute its selected feature vector b through principal geodesic analysis and feature selection. Then we compute its a posteriori probabilities r_f and r_m , the mean vectors \hat{b}_f, \hat{b}_m , and the precision matrices Λ_f, Λ_m . If $r_f > r_m$ then the face is classified as female. Otherwise, the face is classified as male. The posterior probabilities are given by

$$r_{nk} \propto \pi_k |\Lambda_k|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(x_n - \mu_k)^T \Lambda_k (x_n - \mu_k)\right\} . \quad (13)$$

where $k = 1, \dots, K$ is the mixture component, $n = 1, \dots, N$ denotes the data index. Model parameters π_k, μ_k and Λ_k are respectively a priori probability, the mean of selected feature vectors and precision matrices of the k^{th} component. In the variational Bayesian EM (VBEM) algorithm, all of these model parameters are characterized by hyper-parameters, which take into account the uncertainty in the parameter estimation. The parameters r_{nk} are called posteriori probability because they represent the responsibility the k^{th} component takes in explaining the n^{th} observation. The posteriori probability can be arranged into a matrix $R = (r_{nk})$ and will have to satisfy the following conditions:

$$0 \leq r_{nk} \leq 1 . \quad (14)$$

5. Experiments and Comparisons

5.1. 2.5D Facial Needle-maps

The data set used to test the performance of our proposed algorithm is the facial needle-maps extracted from the Max-Planck database of range images, see Fig. 2. It comprises 200 (100 females and 100 males) laser scanned human heads without hair. The facial needle-maps (fields of surface normals) are obtained by first orthographically projecting the facial range scans onto a frontal view plane, aligning the plane according to the eye centers, and then cropping the plane to be 256-by-256 pixels to maintain only the inner part of the face. The surface normals are then obtained by computing the height gradients of the aligned range images. We then apply the principal geodesic analysis (PGA) method to the needle maps to extract features. Whereas PCA can be applied to data residing in a Euclidean space to locate feature subspaces, PGA applies to data constrained to fall on a manifold. Examples of such data are direction fields, and tensor-data.



Figure 2. Examples of Max - Planck needle - maps

5.2. Gender Classification Results

Using the feature selection algorithm outlined above, we first make use of subset-level trace ratio criterion (S-TR) to find the gender discriminating feature subset from the leading 10 PGA features, and compare the results with empirical feature selection. Then, by selecting a different number of features from the leading 30 PGA feature components, we compare the gender classification results from

subset-level trace ratio criterion (S-TR) with those obtained using feature-level feature selection methods, namely a) Maximum-Relevance Minimum-Redundancy (MRMR), b) Mutual Information Based Feature Selection (MIFS). The data used are the PGA feature vectors extracted from the 200 Max-Planck facial needle maps. We randomly select 20 female faces and 20 male faces from the 200 Max-Planck data as test data. The remaining 160 are used as training data.

First, we empirically select the gender discriminating features by examining the distribution of the leading 10 PGA feature components (denoted $b_i, i = 1, \dots, 10$) of the 200 facial needle-maps where the first 100 data to be female, and the following 100 to be male. We normalized the value of PGA feature component by its corresponding eigenvalue and then calculate the normalized mean values of female and male respectively, which are shown in Table. 1. We rank the feature based on the feature-level score, the result is $\{F_1, F_5, F_6, F_2, F_9, F_{10}, F_4, F_8, F_3, F_7\}$

	b_1^{norm}	b_2^{norm}	b_3^{norm}	b_4^{norm}	b_5^{norm}
F	-0.620	-0.252	-0.048	0.085	0.376
M	0.620	0.252	0.048	-0.085	-0.376
	b_6^{norm}	b_7^{norm}	b_8^{norm}	b_9^{norm}	b_{10}^{norm}
F	-0.364	0.031	0.051	0.154	0.092
M	0.364	-0.031	-0.051	-0.154	-0.092

Table 1. Mean values of the leading PGA feature components

We then apply subset-level trace ratio criterion (S-TR) to find the gender discriminating feature subset from the leading 10 PGA feature components, and compare the results with empirical feature selection based on feature-level score. The selected feature subset is $\{F_1, F_6, F_5, F_2, F_9, F_3, F_8, F_7, F_4, F_{10}\}$ which is differs slightly from empirical feature selection.

To explore the data in more detail, we visualize the 1st, 5th, and 6th eigenmodes by showing the mean face together with its deviation along the 1st, 5th and 6th eigenmodes. The visualization is shown in Fig. 3, and by inspection it seems plausible the three eigenmodes do convey some gender information. Turning our attention to the 1st eigenmode, the faces from left to right become more solid in appearance becoming larger and “squarer”, while the cheeks become thinner. These are all masculine characteristics. In the case of 5th component, the faces become more oval and the eyes wider. These are feminine characteristics. In the case of the 6th component, the faces again have more masculine appearance from left to right. Fig. 3 therefore indicates that the gender discriminating features selected using the subset-level trace ratio criterion (S-TR) are at least to some degree consistent with human perception.

For the above experiments, we select the gender discriminating features only from the leading 10 PGA feature

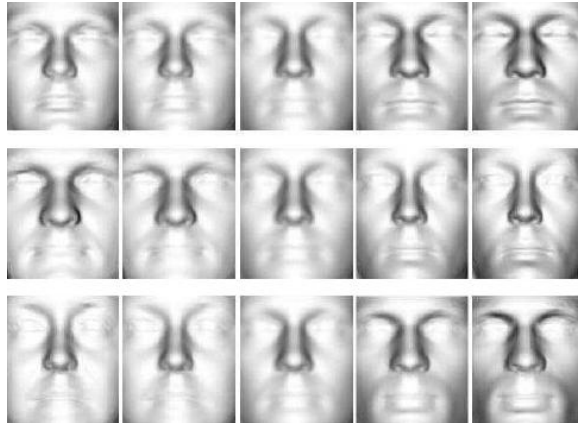


Figure 3. Visualization of the 1st, 5th and 6th eigenmodes. From top to bottom are the 1st, 5th, 6th eigenmodes. The columns are according to the deviation from the mean, from left to right are $\lambda = -30, \lambda = -20, \lambda = 0$ (the mean face), $\lambda = 20$ and $\lambda = 30$,

components. However, on the base of cumulative reason, there are probably additional non leading features that are important for gender discrimination. Therefore, in the following experiments, we extend our attention to the leading 30 features. In Fig. 4, we compare the performance of the three criterion functions. For the selected feature subsets, we apply the variational EM algorithm [3] to fit a mixture of Gaussians model to the d-dimensional feature vectors of the training data set of 160 needle maps. After learning the mixture model, we use a posteriori probability, see Equation(13), to perform gender classification of 40 test faces and we compare the performance of the three criterion functions. At small dimensionality there is little difference between the different methods. However, at higher dimensionality, the features selected by S-TR clearly have a higher discriminability power than the features selected by MRMR and MIFS. The ideal size of the feature subsets for MIFS and MRMR is 4 or 5, where an accuracy 87% and 92.5% respectively are achieved. However, for S-TR we obtain the highest accuracy of 95% when nearly 10 features are selected. This means that with both MRMR and MIFS there is a tendency to overestimate the redundancy between features, since they neglect the possible feature combinations. As a result some important features can be discarded, which in turn leads to information loss.

Our proposed method therefore leads to an increase in performance at high dimensionality. By considering higher order dependencies between features, we avoid premature rejection on the basis of redundancy, a well documented problem known to exist in MRMR and MIFS [11]. This gives feature sets which are more informative to our classification problem.

We illustrate the learning process and the classification results using the selected feature subset. The learning pro-

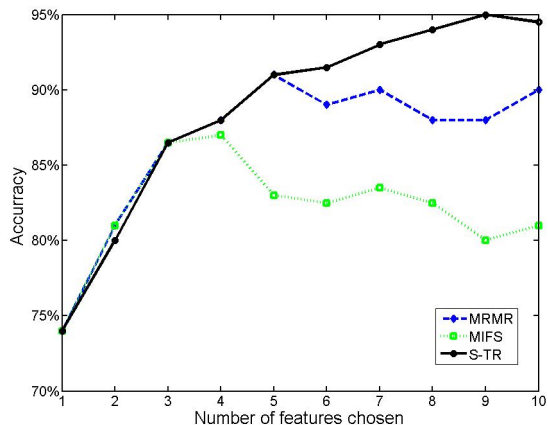


Figure 4. Average classification accuracies: S-TR shows significant benefit compared to criteria of MRMR and MIFS

cess of the EM step in the VBEM algorithm is visualized in Fig. 5 by showing the models learned, a) on initialization, b) on convergence. Each stage is visualized by showing the distribution of the 1st and 6th PGA feature components as a scatter plot. From Fig. 5, it is clear that on convergence of the VBEM algorithm (after 50 iterations), the data are well clustered according to gender. We see that after convergence, only two components survive. In other words, there is an automatic trade-off in the Bayesian model between fitting the data and the complexity of the model, in which the complexity penalty arises from components whose parameters are forced away from their prior values.

The gender classification accuracy achieved using the selected feature subset obtained by S-TR method is 95%. The classification results are shown in the Fig. 6, where the first 20 faces are female and the remaining 20 are male. The misclassified faces (the facial needle-maps) are visualized on the top. These two are faces are females misclassified as males and all male faces are correctly classified. It is clear from the figure that adequate gender separability provided by the selected feature subset, as the unsupervised nature of the VBEM algorithm and Gaussian mixture models. In addition, the misclassified faces from the mixture model classifier are at least to some degree ambiguous for human judgement too.

6. Conclusions

This paper presents subset-level trace ratio criterion (S-TR) for feature selection. The proposed method offers two major advantages. First, the S-TR criteria takes into account feature interactions which selecting features based on subset-level, overcoming the problem of overestimated redundancy. As a result the features associated with the greatest amount of joint information can be preserved. Second, the variational EM algorithm and a Gaussian mixture

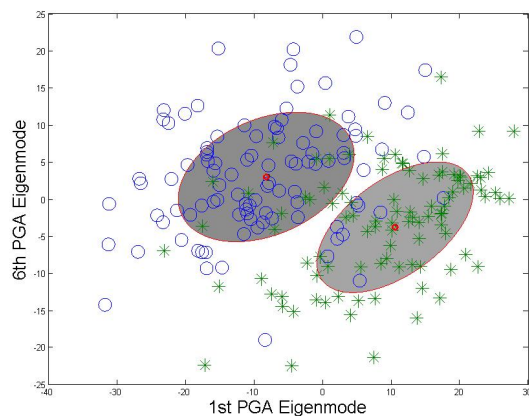
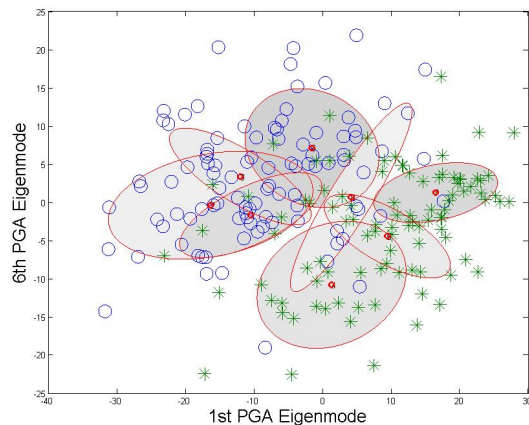


Figure 5. VBEM of $K = 8$ learning process visualized on 1st and 6th PGA feature components, in which the ellipses denote the one standard-deviation density contours for each components, and the density of red ink inside each ellipse corresponds to the mean value of the mixing coefficient for each component. The initial number of component is $K = 8$, during the learning process, some components whose expected mixing coefficient are numerically indistinguishable from zero are not plotted

model are applied to the selected feature subset improving the overall classification.

References

- [1] K. Balagani, V. Phoha, S. Iyengar, and N. Balakrishnan. On Guo and Nixon's Criterion for Feature Subset Selection: Assumptions, Implications, and Alternative Options. *IEEE TSMC-A: Systems and Humans*, 40(3):651–655, 2010. 2
- [2] R. Battiti. Using Mutual Information for Selecting Features in Supervised Neural Net Learning. *IEEE Transactions on Neural Networks*, 5(4):537–550, 2002. 1
- [3] C. Bishop. *Pattern Recognition and Machine Learning*, volume 4. Springer New York, 2006. 4, 5
- [4] S. Buchala, N. Davey, T. Gale, and R. Frank. Principal Component Analysis of Gender, Ethnicity, Age, and Identity of Face Images. *Proc. IEEE ICMI*, 2005. 1, 2

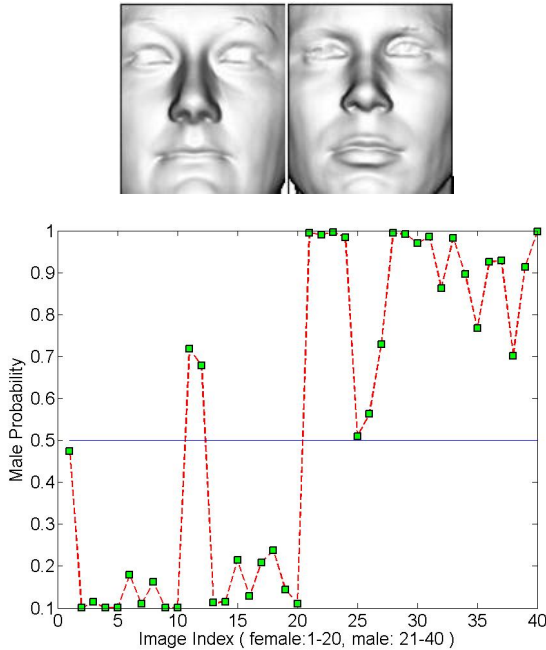


Figure 6. Classification result based on the selected feature subset from S-TR method. The top is the misclassified female faces, the bottom is the classification result calculated by posteriori probability

- [5] Y. Chang, C. Hu, and M. Turk. Manifold of facial expression. In *IEEE International Workshop on Analysis and Modeling of Faces and Gestures.*, pages 28–35. IEEE, 2003. 1
- [6] H. Cheng, Z. Qin, W. Qian, and W. Liu. Conditional Mutual Information Based Feature Selection. In *IEEE International Symposium on Knowledge Acquisition and Modeling*, pages 103–107, 2008. 2
- [7] T. Cover. The Best Two Independent Measurements Are Not The Two Best. *IEEE TSMC*, 4(1):116–117, 2010. 1
- [8] T. Cover, J. Thomas, and J. Wiley. *Elements of Information Theory*, volume 1. Wiley-Interscience, 1991. 1
- [9] P. Fletcher, C. Lu, S. Pizer, and S. Joshi. Principal Geodesic Analysis for The Study of Nonlinear Statistics of Shape. *IEEE Transactions on Medical Imaging.*, 23(8):995–1005, 2004. 2
- [10] A. Graf and F. Wichmann. Gender classification of human faces. In *Biologically Motivated Computer Vision*, pages 1–18. Springer, 2010. 1
- [11] M. Gurban and J. Thiran. Information Theoretic Feature Extraction for Audio-visual Speech Recognition. *IEEE Transactions on Signal Processing*, 57(12):4765–4776, 2009. 5
- [12] N. Kwak and C. Choi. Input Feature Selection by Mutual Information Based on Parzen Window. *IEEE TPAMI*, 24(12):1667–1671, 2002. 2
- [13] K. Lee, J. Ho, M. Yang, and D. Kriegman. Video-based face recognition using probabilistic appearance manifolds. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 313–320, 2003. 1
- [14] F. Nie, S. Xiang, Y. Jia, C. Zhang, and S. Yan. Trace ratio criterion for feature selection. In *Proceedings of the 23rd national conference on Artificial intelligence-Volume 2*, pages 671–676. AAAI Press, 2008. 1, 3
- [15] H. Peng, F. Long, and C. Ding. Feature Selection Based on Mutual Information: Criteria of Max-dependency, Max-relevance, and Min-redundancy. *IEEE TPAMI*, 27(8):1226–1238, 2005. 2
- [16] W. Smith and E. Hancock. Facial shape-from-shading and recognition using principal geodesic analysis and robust statistics. *International Journal of Computer Vision*, 76(1):71–91, 2008. 1
- [17] M. Turk and A. Pentland. Face recognition using eigenfaces. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 586–591. IEEE, 1991. 1
- [18] J. Wu, W. Smith, and E. Hancock. Gender classification using shape from shading. In *British machine vision conference*. Citeseer, 2007. 1
- [19] H. Yang and J. Moody. Feature Selection Based on Joint Mutual Information. In *Proceedings of International ICSC Symposium on Advances in Intelligent Data Analysis*, pages 22–25, 1999. 2