Learning a subspace for face image clustering via trace ratio criterion

Chenping Hou,[a,b] Feiping Nie,[b] Changshui Zhang,[b] and Yi Wu[a]
[a]National University of Defense Technology, Department of Mathematics and Systems Science, Changsha,
Hunan, 410073, China
[b]Tsinghua University, Department of Automation, Beijing,
10084, China
E-mail: hcpnudt@hotmail.com

Abstract. Face clustering is gaining ever-increasing attention due to its importance in optical image processing. Because traditional clustering methods do not specify the particular characters of the face image, they are not suitable for face image clustering. We propose a novel approach that employs the trace ratio criterion and specifies that the face images should be spatially smooth. The graph regularization technique is also applied to constrain that nearby images have similar cluster indicators. We alternately learn the optimal subspace and the clusters. Experimental results demonstrate that the proposed approach performs better than other learning methods for face image clustering.
[DOI: 10.1117/1.3149850]

Subject terms: face image clustering; subspace learning.

Paper 080957LR received Dec. 9, 2008; revised manuscript received Apr. 1, 2009; accepted for publication Apr. 20, 2009; published online Jun. 5, 2009.

1 Introduction

Human face clustering is an important research direction in the field of optical image processing. It has been successfully used in many fields. For example, in face image retrieval, if we can automatically cluster different kinds of face images, then it will beneficial to discover the latent similarities among various face images. The computational cost of searching for an interested face image will also reduce considerably.

Currently, a lot of research has been proposed for face image processing. However, most of it focuses on learning a subspace for face classification (i.e., classify the face images in the learned subspace with a certain number of labeled samples). The most prominent approach is called the *Fisherface*.[1] It employs the supervised method, *linear discriminant analysis* (LDA),[2] for face classification.

There is little work dedicated to learning a subspace for face image clustering. One direct way is to employ the traditional clustering techniques (e.g., K-means[3]) for clustering. Nevertheless, it is difficult for K-means to achieve a satisfied accuracy because the vector representations of face images have very high dimensionality (commonly, the number of pixels of a image). Then, one of the most famous works, which is named PCA+K-means, is proposed. It employs the unsupervised *principle component analysis* (PCA)[2] technique to compute the famous *Eigenface*[1] and then applies the K-means approach for clustering. Although it has achieved prominent performances in many applications, PCA+K-means seems unsuitable for face image clustering. It takes into account few considerations about the special characters and manifold structures of face images.[4] Recently, Ding et al. extended the traditional supervised LDA method for clustering.[5] The method, which is called LDA-Km, combines LDA with K-means and learns the subspace and clusters alternately. It has been reported that LDA-Km performs better than the K-means and the PCA+K-mean approaches.[5] However, it also gives no considerations to the manifold structure and the spatial smooth character[6] of face images. Additionally, LDA-Km applies the traditional ratio trace criterion, not the more prominent trace ratio criterion.[7] The accuracy of LDA-Km is not as high as expected for face image clustering.

In this paper, we propose a new method, which is named as *subspace clustering via trace ratio* (SCTR) for face image clustering. First, it applies the trace ratio criterion, which has been shown more effective for discriminative subspace learning.[7] Second, we employ the spatial smooth regularizer, which represents the particular character of images. Finally, we consider the manifold structure by intentionally adding the graph-smooth constraint. After computing the optimal clustering and learning the subspace alternately, we can finally derive the subspace and clustering results simultaneously. Experiments show that SCTR performs better than the state-of-the-art approaches.

2 Leaning a Subspace for Face Image Clustering

2.1 *Problem Formulation*

Assume $X=[x_1,x_2,\ldots,x_l]$ are vector representations of l face images, which are all of $D_1\times D_2$ resolutions. Here $x_i\in\mathbb{R}^D$ for $i=1,2,\ldots,N$ and $D=D_1\times D_2$. The number of clusters is predefined as C. The purpose of our method is to find the subspace in which these face images can be optimally clustered. More concretely, under a given criterion, we plan to find the transformation matrix $W\in\mathbb{R}^{D\times d}$ and simultaneously, the indicator matrix $F\in\mathbb{R}^{l\times C}$. Here, d is the dimensionality of the subspace. $F_{ij}=1/\sqrt{l_j}$ if x_i belong to the jth cluster, and $F_{ij}=0$ otherwise, where l_j is the number of samples in the jth cluster.

It has been shown[7] that the trace ratio criterion is much more effective than the ratio trace, which is widely used in the traditional LDA approach and its variants. We employ the trace ratio criterion in our approach. Mathematically, assume $Y=[y_1,y_2,\ldots,y_l]$ are the embddings of X (i.e., $Y=W^TX$). The within scatter matrix S_w and the total-scatter matrix S_t are defined as follows:

$$S_{\mathrm{w}}=\sum_{j=1}^{C}\sum_{i=1}^{l_j}(y_i-m_j)(y_i-m_j)^T=YL_{\mathrm{w}}Y^T,$$

$$S_{\mathrm{t}}=\sum_{i=1}^{l}(y_i-m)(y_i-m)^T=YL_{\mathrm{t}}Y^T, \tag{1}$$

where $m_j=(1/l_j)\Sigma_{i=1}^{l_j}y_i$ $(j=1,2,\ldots,C)$ is the mean of the samples in the j'th cluster, $m=(1/l)\Sigma_{i=1}^{l}y_i$ is the mean of all samples. The corresponding L_{w} and L_{t} are

0091-3286/2009/$25.00 © 2009 SPIE

$$L_{\mathrm{w}} = I - FF^T, \quad L_{\mathrm{t}} = I - \frac{1}{l} ee^T, \tag{2}$$

where e is an l-dimensional column vector with all ones. The trace ratio criterion proposed in Ref. 7 is

$$\arg \min_{W^T W = I} \frac{\mathrm{tr}(W^T S_{\mathrm{w}} W)}{\mathrm{tr}(W^T S_{\mathrm{t}} W)}. \tag{3}$$

We now consider the spatial character of images. Because an image represented in the plane is intrinsically a matrix. The pixels spatially close to each other may be correlated. Although we have $D_1 \times D_2$ pixels per image, this spatial correlation suggests the real number of freedom is far less. We employ a spatial regularizer (i.e., the Laplacian penalty) to constrain the coefficients to be spatially smooth.[6] In brief, we define the $D_j \times D_j$ $(j=1,2)$ matrices that yield the discrete approximation of the second-order derivation as follows:

$$M_j = \frac{1}{h_j^2} \begin{pmatrix} -1 & 1 & & & 0 \\ 1 & -2 & 1 & & \\ & \cdots & \cdots & \cdots & \\ & & 1 & -2 & 1 \\ 0 & & & 1 & -1 \end{pmatrix}. \tag{4}$$

Here, $h_j = 1/D_j$ for $j=1,2$. The discrete approximation for the two-dimensional Laplacian is a $D \times D$ matrix

$$\Delta = M_1 \otimes I_2 + I_1 \otimes M_2, \tag{5}$$

where I_j is the $D_j \times D_j$ identity matrix for $j=1,2$. $\otimes$ represents the Kronecker product of two matrixes.

For a $D_1 \times D_2$ image vector x, $\|\Delta x\|^2$ is proportional to the sum of squared differences between nearby grid points in that image. It can be used to measure the smoothness of an image on a $D_1 \times D_2$ lattice. Because each column vector of W can be regarded as a basis image, we add $\alpha \Delta^T \Delta$ to S_{w} where $0 < \alpha < 1$ is a balance parameter that controls the smoothness of the estimator.

Finally, as in most learning-based face image processing approaches[1] we assume that the face images, which belong to the same cluster, are nearly lying on a low dimensional manifold. The graph Laplacian[4] (i.e., L) is employed to represent this character. Intuitively, the cluster indicators of face images belonging to the same cluster should be identical. In other words, if we regard that the cluster indicators are the output of a function defined on all face images, this intuition indicates that the predefined function should be smooth on these manifolds. Mathematically, it is equal to minimize $\mathrm{tr}(F^T L F)$.

In summary, we define SCTR as the solution to the following problem:

$$\begin{aligned} &\arg \min_{W^T W = I, F^T F = I} (1-\beta) \frac{\mathrm{tr}[(1-\alpha) W^T S_{\mathrm{w}} W + \alpha W^T \Delta^T \Delta W]}{\mathrm{tr}(W^T S_t W)} \\ &\quad + \beta\, \mathrm{tr}(F^T L F) = \arg \min_{W^T W = I, F^T F = I} (1 \\ &\quad - \beta) \frac{\mathrm{tr}\{(1-\alpha) W^T [X(I - FF^T) X^T + \alpha \Delta^T \Delta] W\}}{\mathrm{tr}(W^T X L_{\mathrm{t}} X^T W)} \\ &\quad + \beta\, \mathrm{tr}(F^T L F), \end{aligned} \tag{6}$$

where $0 < \beta < 1$ is also a balance parameter.

2.2 *Solution*

There are totally two different kinds of variables that should be optimized (i.e., W and F). It is difficult to compute them simultaneously. We alternately optimize them.

2.2.1 *Stage one: fixing F and optimizing W*

In this situation, the optimization problem in Eq. (6) becomes

$$\arg \min_{W^T W = I} \frac{\mathrm{tr}\{W^T [(1-\alpha) X(I - FF^T) X^T + \alpha \Delta^T \Delta] W\}}{\mathrm{tr}(W^T X L_{\mathrm{t}} X^T W)}. \tag{7}$$

It has the same form as the problem proposed in Ref. 7, except for a little differences in the formulation of the numerator. Thus, we can directly employ the same technique. It is an iterative procedure. For our problem, in the p'th step, we solve a trace difference problem

$$\begin{aligned} &\arg \min_{W^T W = I} \mathrm{tr}\{W^T [(1-\alpha) X(I - FF^T) X^T + \alpha \Delta^T \Delta \\ &- \lambda^p X L_{\mathrm{t}} X^T] W\}, \end{aligned}$$

where λ^p is the trace ratio value calculated from the previous projection matrix W^{p-1} in the previous step. Please see Ref. 7 for more details.

2.2.2 *Stage two: fixing W and optimizing F*

In this case, because L_{t} and Δ have no relation with F, the optimization problem in Eq. (6) is equivalent to

$$\begin{aligned} &\arg \max_{F^T F = I} \mathrm{tr}\{F^T [(1-\beta)(1-\alpha) X^T W W^T X \\ &- \beta\, \mathrm{tr}(W^T X L_{\mathrm{t}} X^T W) L] F\}. \end{aligned} \tag{8}$$

This problem can be effectively solved by spectral decomposition technique. In fact, the optimal F to the problem in Eq. (8) is formed by the eigenvectors corresponding to the C largest eigenvalues of $(1-\beta)(1-\alpha) X^T W W^T X - \beta\, \mathrm{tr}(W^T X L_{\mathrm{t}} X^T W) L$.

2.3 *Discussions*

There are two essential parameters (i.e., $0 < \alpha < 1$ and $0 < \beta < 1$) in our method. They control the smoothness of estimator and the balance of two objectives. When $\alpha = 0$, SCTR totally ignores the spatial character of the face image. When $\alpha \rightarrow 1$, SCTR tends to only choose a spatial smoothest basis and ignore the discriminative information. When $\beta = 0$, we ignore the smooth constraint on F. If $\beta = 1$, SCTR only concerns the smoothness of F.

Parameter determination is an essential task in most of the learning algorithms.[6] Among various kinds of methods, grid search is probably the simplest and most widely used one for unsupervised learning. Because we have constrained $0 < \alpha < 1$ and $0 < \beta < 1$, we apply the grid search technique in this paper. More concretely, we set α and β by searching the grid $\{0.1, 0.2, \ldots, 0.9\} \times \{0.1, 0.2, \ldots, 0.9\}$.

There are another three problems that should be indicated here. (i) Because SCTR is an iterative approach, a

Table 1 Mean accuracy (Acc) results of different methods on four face image data sets.

	K-means	PCA+K-means	LDA-Km	SCTR
Yale	0.379	0.385	0.459	**0.491**
ORL	0.497	0.513	0.607	**0.674**
Umist	0.381	0.383	0.468	**0.501**
Feret	0.374	0.372	0.526	**0.551**

direct way to initialize the iteration is to first compute F by K-means and then optimize the problems in Eqs. (7) and (8) alternately. (*ii*) The optimal F that maximizes Eq. (8) may not have the form of the indicator matrix that we specified in the previous part. We apply the discretization procedure in Ref. 8 to solve this problem. (*iii*) Because we have derived W and F, it is easy to compute the embedding and the cluster of a new image by using W and F directly.

3 Experiments

We employ four different kinds of face image data sets for illustration. They are the Yale data, the ORL data, the Umist data, and the first 20 classes of Feret data. All these images are rescaled to 16×16 resolutions.

We compare SCTR with K-means, PCA+K-means, and LDA-Km. Two standard measurements—the accuracy (Acc) and normalized mutual information (NMI)—are used. The parameters α and β are determined by grid search and $d=C-1$. We randomly initialize K-means and repeat for 100 times. The average values of these Accs are listed in Table 1. Figure 1 shows the corresponding mean of these NMIs.

As seen from Table 1 and Fig. 1, SCTR performs the best. It achieves both the highest mean Acc and the largest mean NMI in all cases. Because all the images are rescaled to be low resolution and the dimensionality is not so high, it seems that K-means and PCA+K-means have the similar performances on these data sets. LDA-Km performs better than K-means and PCA+K-means.

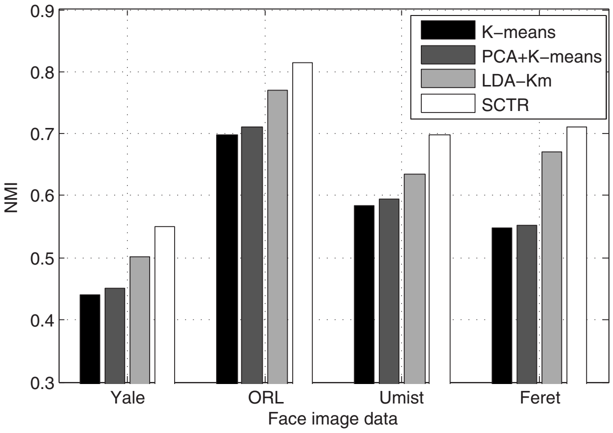

Fig. 1 Mean normalized mutual information (NMI) on the Yale, the ORL, the Umist and the Feret data sets.

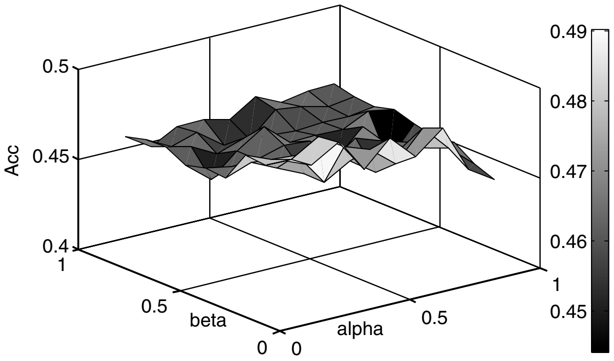

Fig. 2 Acc versus α and β on Yale by grid search.

We have also done some experiments on the Yale data set with different parameters, which are set by the grid search. The Acc results versus different α and β are shown in Fig. 2. It seems that for Yale, smaller α and β are more suitable. The best parameter $\alpha=0.3$ and $\beta=0.1$.

4 Conclusions

We propose a new subspace learning method for face image clustering. It integrates the spatial characters and the manifold structures of face images. By alternately computing the subspace and clusters, it performs best among all the involved approaches.

Acknowledgments

This work is partly supported by NSF China (Grants No. 60673090 and No. 60721003), the Hunan Provincial Innovation Foundation for Postgraduate and the Graduate School of National University of Defense Technology.

References

1. P. Belhumeur, J. Hepanha, and D. Kriegman, "Eigenfaces vs. fisherfaces: recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.* **19**(7), 711–720 (1997).
2. R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. Wiley-Interscience, Hoboken, NJ (2000).
3. L. Parsons, E. Haque, and H. Liu, "Subspace clustering for high dimensional data: a review," *SIGKDD Explorations* **6**, 90–105 (2004).
4. M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: a geometric framework for learning from examples," *J. Mach. Learn. Res.* **7**, 2399–2434 (2006).
5. C. Ding and T. Li, "Adaptive dimension reduction using discriminant analysis and K-means clustering," *Proc. 24th Int. Conf. on Machine Learning* pp. 521–528, ACM Press, Madison, WI (2007).
6. T. Hastie, A. Buja, and R. Tibshirani, "Penalized discriminant analysis," *Ann. Stat.* **23**, 73–102 (1995).
7. H. Wang, S. Yan, D. Xu, X. Tang, and T. Huang, "Trace ratio vs. ratio trace for dimensionality reduction," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2007)*, June 2007, IEEE Computer Society, Washington, DC (2007).
8. S. Yu and J. Shi, "Multiclass spectral clustering," *IEEE International Conference on Computer Vision (ICCV 2003)*, pp. 313–319, October 2003, IEEE Computer Society, Washington, DC (2003).