Pursuing Informative Projection on Grassmann Manifold

Dahua Lin, Shuicheng Yan
Dept. of Information Engineering
The Chinese University of Hong Kong
{dhlin4, scyan}@ie.cuhk.edu.hk

Xiaoou Tang
Microsoft Research Asia
Beijing, China
xitang@microsoft.com

Abstract

Inspired by the underlying relationship between classification capability and the mutual information, in this paper, we first establish a quantitative model to describe the information transmission process from feature extraction to final classification and identify the critical channel in this propagation path, and then propose a Maximum Effective Information Criteria for pursuing the optimal subspace in the sense of preserving maximum information that can be conveyed to final decision. Considering the orthogonality and rotation invariance properties of the solution space, we present a Conjugate Gradient method constrained on a Grassmann manifold to exploit the geometric traits of the solution space for enhancing the efficiency of optimization. Comprehensive experiments demonstrate that the framework integrating the Maximum Effective Information Criteria and Grassmann manifold-based optimization method significantly improves the classification performance.

1 Introduction

Despite the great advance of face recognition techniques in past decades, it remains one of the most challenging topics in computer vision. The crucial obstacle hindering the improvement of recognition performance is "the curse of dimensionality". To mitigate the difficulties incurred by high dimension, a variety of subspace analysis methods are developed to reduce the representation dimension. Representative subspace analysis algorithms include Principal Component Analysis (PCA) [7] and Linear Discriminant Analysis (LDA) [10].

PCA [7] finds a subspace preserving most of the variations and decorrelating the components in the principal subspace. Though simple and effective in dimension reduction, with the goal of best reconstruction, PCA is not necessarily optimal to classification. LDA [10] and its variations [3] [15] aim at acquiring a subspace well separating the samples of different classes by maximizing the trace-ratio of between-class scattering matrix to within-class scattering matrix. Though it is advocated that LDA effectively extracts the discriminative information, however, in the formulation of LDA, the so-called "discriminative information" is only a notion without clear concept and quantitative description.

Outside traditional statistical learning, information theory provides us with an interestingly new perspective for viewing the classification problem. Intuitively speaking, the more information we know about the classes from samples, the better we can classify the new samples. This rationale is also supported by some theoretical analysis [5] [14] in information theory. Recently some related works have been done to apply information theory for supervised learning. Vidal-Naquet. et al proposed "Informative feature selection" [8] to select features by maximizing mutual information with greedy search; Liu. et al proposed "Kullback-Leibler Analysis" [4] to sequentially learn a set of linear features for face detection; Wu et. al [16] advocated to employ "Balanced information gain" for feature selection; and Vasconcelos [14] discussed the "Maximum Margin Diversity" algorithm employing an information-based ranking strategy. All of them use an information theoretical criteria for sequential feature selection.

However, there are two main limitations for these methods. First, although these works realized the importance of information for classification, however, they fail to offer an insight into how the information is utilized by classification, which is indeed a crucial aspect affecting performance as shown later. Second, these methods are mainly used for sequential feature selection based on greedy search strategy or ranking strategy; they fail to obtain an optimal multidimensional transform of features like that in algorithms such as LDA. The obstacle impeding applying information principle for feature extraction consists in the computational difficulties in optimization due to the special form of information theory-based objective functions, where the probability density function is explicitly involved and affected by the variables being optimized. Recently, Torkkola [13] proposed a method for feature extraction by maximizing mutual information. However, it suffers several drawbacks: 1)

Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)
0-7695-2597-0/06 $20.00 © 2006 IEEE

To obtain a tractable mathematical expression, its formulation uses Renyi entropy instead of Shannon entropy; while such a replacement is not well justified, and what effect it would bring is not clear; 2) The optimization is based on a multidimensional density estimation, thus it may be inaccurate and not robust; 3) Its effectiveness is not sufficiently validated in real data.

In this paper, we propose a novel information-theoretical method to address the above issues. Based on the theoretical relation between information theory and classification, we first establish a channel model quantitatively describing the classification procedure as an information transmission process where the information is delivered through cascaded channels. By analyzing the information propagation, we found that the classification related information is lost when it goes through a deterministic transforms. And the effective information that can be finally conveyed to decision stage is determined by the distribution of metric values, which serves as a pivot in the information flow. Based on this rationale, we derive a maximum effective information criteria for optimizing the projection. Exploiting the fact that the distribution of metrics is in 1D space, we effectually overcome the computational hurdle in applying information theory to learning feature projection. Observing the orthogonality constraint of the projection matrix and the rotation invariance of objective function, we employ the Grassmann manifold to guide the optimization process, which is a high-dimensional manifold consisting of all projection matrices under homogeneity condition. Then the complicated constrained optimization problem is converted to an unconstrained problem on a curved hypersurface with much lower degree of freedom, which follows a much more efficient procedure. Extensive experiments convincingly support our theoretical analysis and demonstrate the superiority of our approach over traditional subspace algorithms.

2 Maximizing Effective Information

2.1 Relation between Mutual Information and Classification

We first briefly review two fundamental concepts in information theory: *entropy* and *mutual information*, which is for measuring the uncertainty of random variables and the mutual information conveyed between two random variables respectively. In the following text, we denote the entropy of variable x by $H(x)$ and the mutual information between variable x and y by $I(x;y)$.

Fano established a lower bound on classification errors in terms of class posterior entropy in the well-known Fano's inequality [5], and it has been realized that there exists intrinsic relationship between information theory and pattern recognition. The study by Vasconcelos [14] further shows that "*infomax solutions are near-optimal in minimum Bayes error sense*", which gives strong theoretical support to applying maximum information criteria in classification.

2.2 Linear Projection and Metric

In pattern recognition, each face is represented as a d-dimensional vector $\mathbf{x} = (x_1, x_2, \ldots, x_d)^T$. In supervised learning approach, a model is learned from a training set $\{(\mathbf{x}_i, c_i)\}_{i=1}^n$. Here n is the number of samples, $\mathbf{x}_i$ is the vector representation of the i-th sample, while $c_i = l(\mathbf{x}_i)$ is the corresponding class label.

A linear projection is characterized by an orthogonal $d \times p$ matrix $\mathbf{A}$. Here, p is the dimension of the subspace. For a vector $\mathbf{x}$, it can be projected to the subspace by $\mathbf{y} = \mathbf{A}^T\mathbf{x}$. It is typical to measure the similarity in terms of Euclidean distance in the projected space:

$$\begin{aligned} s(\mathbf{x}_1, \mathbf{x}_2|\mathbf{A}) &= ||\mathbf{A}^T\mathbf{x}_1 - \mathbf{A}^T\mathbf{x}_2||^2 \\ &= (\mathbf{x}_1 - \mathbf{x}_2)^T\mathbf{A}\mathbf{A}^T(\mathbf{x}_1 - \mathbf{x}_2). \end{aligned} \quad (1)$$

Due to the large number of classes in pattern recognition, it is expensive and unreliable to build statistical model for each individual class. To address the complexity of the multi-class classification problem, we convert the problem into a two-class one using a simple strategy similar to Bayesian face [9].

¿From Eq.(1), we can see that in a metric space with L^2 distance, the metric depends solely on the difference vector $\mathbf{u} = \mathbf{x}_1 - \mathbf{x}_2$ between two samples but not on their absolute values. Therefore, according to whether two samples are from the same class, we categorize their difference vectors into two types: *intra-class difference* and *extra-class difference*; and the corresponding sample vector sets are denoted as $\Omega_I = \{\mathbf{u} = \mathbf{x}_1 - \mathbf{x}_2 | l(\mathbf{x}_1) = l(\mathbf{x}_2)\}$ and $\Omega_E = \{\mathbf{u} = \mathbf{x}_1 - \mathbf{x}_2 | l(\mathbf{x}_1) \neq l(\mathbf{x}_2)\}$. Consequently, the multi-class problem is converted into a two-class one, and the metric is computed based on $\mathbf{u}$ as

$$s(\mathbf{u}|\mathbf{A}) = \mathbf{u}^T\mathbf{A}\mathbf{A}^T\mathbf{u}. \quad (2)$$

2.3 Channel Model and Effective Information

2.3.1 Information Channel Model

To gain an insight into the information transmission process, we first review the procedure in metric-based classification as illustrated in Figure 1. The whole procedure is divided into three stages: the transform stage which projects the source difference vector into a low dimensional subspace, the metric evaluation stage producing the metric value, and the decision stage making the judgment. In the

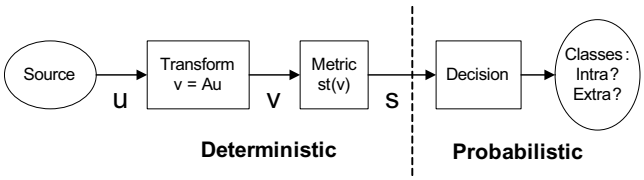

Figure 1. Information Channel Model.

presented channel model, each stage is modelled as an information channel with information propagated from input variable to output variable.

¿From Figure 1, we can see that, when the transform matrix $\mathbf{A}$ is given, the first two stages are both deterministic channels, *i.e.* when input variable is given, the output is determined without any uncertainty; while the last channel is a probabilistic channel, *i.e.* when metric is given, there still exists some uncertainty on final judgment. The pivot connecting the deterministic channels and probabilistic channel is the metric value. In the whole process, information from sample features are first centralized to form the metric values and then the information embedded in the metric values is utilized to support the final decision.

2.3.2 Maximum Effective Information Criteria

Motivated by the rationale, we further analyze the channel model in order to quantitatively describe the process of information flow.

Lemma *Suppose we have three random variables x, y, z, information is propagated through a deterministic channel from x to y, and then through a probabilistic channel from y to z. Then we have*

$$I(x;z) = I(x,y;z) = I(y;z) + I(x;z|y). \tag{3}$$

The proof of Eq. (3) is given in Appendix A.

Eq.(3) reveals two important points: (1) the deterministic process will not add new information; (2) in the information channel model, the information propagated from x to z consists of two parts: one is *finally conveyed to* z by y, which is equal to $I(z;y)$, and the other $I(z,x|y)$ is *lost in the deterministic transform*.

As to the classification process, the information is transferred through the following path: $\mathbf{u} \Rightarrow (\mathbf{v} = \mathbf{A}\mathbf{u}) \Rightarrow (s = s_t(\mathbf{v}) = s(\mathbf{u}|\mathbf{A})) \Rightarrow (c)$. Here, $c \in \{0,1\}$ represents whether the vector is intra-class or extra-class. For a given sample space, the sample distribution is fixed (though unknown), thus the mutual information $I(\mathbf{u};c)$ depending only on class-conditional distributions is also fixed. According to Eq. (3), we have

$$I(\mathbf{u};c) = I(s;c) + I(\mathbf{u};c|s). \tag{4}$$

As analyzed above, among the total information $I(\mathbf{u};c)$, only the part $I(s;c)$, which we call *Effective Information*, contributes to the final classification, while $I(\mathbf{u};c|s)$ is lost in the procedure of generating the metric value. Since the effective mutual information is a function of $\mathbf{A}$, we denote it as $I(s;c|\mathbf{A})$; and the optimal $\mathbf{A}$ should optimize the *Maximum Effective Information* as follows

$$\hat{\mathbf{A}} = \underset{\mathbf{A}}{\operatorname{argmax}}\, I(s;c|\mathbf{A}), \tag{5}$$

here

$$I(s;c|\mathbf{A}) = H(s|\mathbf{A}) - H(s|c;\mathbf{A}). \tag{6}$$

Denote $p_s(s|\mathbf{A})$ as probability density function (pdf) of metrics of all difference vectors, $p_s^{(I)}(s;\mathbf{A})$ as pdf of metrics of intra-class differences, $p_s^{(E)}(s;\mathbf{A})$ as pdf of metrics of extra-class differences, then we have

$$H(s|\mathbf{A}) = -\int p_s(s|\mathbf{A}) \log(p_s(s|\mathbf{A}))ds, \tag{7}$$

$$\begin{aligned} H(s|c;\mathbf{A}) = &-P(\Omega_I)\int p_s^{(I)}(s;\mathbf{A}) \log(p_s^{(I)}(s;\mathbf{A}))ds \\ &-P(\Omega_E)\int p_s^{(E)}(s;\mathbf{A}) \log(p_s^{(E)}(s;\mathbf{A}))ds. \end{aligned} \tag{8}$$

2.3.3 Criteria Discussions

Here, we discuss some aspects of the Maximum Effective Information Criteria as follows:

First, conventional information theoretical approaches directly concern $I(\mathbf{v};c)$, while our method maximizes $I(s;c)$. That is because the information transmission from $\mathbf{v}$ and c is not totally free, but restricted by the classification procedure, which inevitably leads to information lost during metric evaluation. Thus the actual amount of information contributing to final classification is $I(s;c)$, so it is more justifiable to maximize $I(s;c)$ instead of $I(\mathbf{v};c)$.

Second, we can further show that in our formulation,

$$I(s;c) = P(\Omega_I)\mathrm{KL}(p(s|\Omega_I)||p(s)) + P(\Omega_E)\mathrm{KL}(p(s|\Omega_E)||p(s))$$

where $KL(p||q) = \int_s p\log(p/q)ds$ is the Kullback Leibler divergence between distributions p and q, reflecting the discrepancy between the two distributions. Hence, $I(s;c)$ actually embodies the differences between class conditional distribution and average distribution. A large information means a large difference between intra-person and extra-person distributions to average distribution, and hence they can be better distinguished.

Third, to optimize the objection function, we only need to estimate the 1-D distribution of the metric values, which can be effectively accomplished in a nonparametric manner. Hence, the merits brought by our formulation are twofold: on one hand, the difficulty of modelling the high-dimensional distribution is inherently eliminated; on the

other hand, by directly estimating the distribution of metrics, our model does not rely on any parametric distribution assumption such as Gaussian.

2.4 Parzen Window Approximation

For 1D distribution with n observed samples denoted as $\{x_i\}|_{i=1}^n$, we can approximate its pdf by Parzen window[12] as

$$\mathrm{pz}(x;\{x_i\}|_{i=1}^n) = \frac{1}{nh}\sum_{i=1}^{n}\phi\left(\frac{x-x_i}{h}\right), \qquad (9)$$

here pz is a function of x, parameterized by a set of known samples $\{x_i\}|_{i=1}^n$, and $\phi(x) = \exp(-x^2)/\sqrt{2\pi}$.

Suppose the prior probability of intra-class difference vectors and extra-class difference vectors are P_I and P_E respectively. The difference vectors available in training set are categorized into two sets: n_I intrapersonal difference vectors: $\mathbf{u}_1^{(I)}, \mathbf{u}_2^{(I)}, \ldots, \mathbf{u}_{n_I}^{(I)}$ and n_E extrapersonal difference vectors: $\mathbf{u}_1^{(E)}, \mathbf{u}_2^{(E)}, \ldots, \mathbf{u}_{n_E}^{(E)}$. Their corresponding metric values are denoted as $s_1^{(I)}, s_2^{(I)}, \ldots, s_{n_I}^{(I)}$ and $s_1^{(E)}, s_2^{(E)}, \ldots, s_{n_E}^{(E)}$. Then we can approximate the probability density functions as

$$p_s^{(I)}(s) = \mathrm{pz}(s;\{s_i^{(I)}\}|_{i=1}^{n_I}), \qquad (10)$$

$$p_s^{(E)}(s) = \mathrm{pz}(s;\{s_i^{(E)}\}|_{i=1}^{n_E}), \qquad (11)$$

$$p_s(s) = P_I p_s^{(I)}(s) + P_E p_s^{(E)}(s). \qquad (12)$$

Based on the Large Number Law, we can approximate the mathematical expectation by sample mean as

$$H(s|\mathbf{A}) = -\frac{P_I}{n_I}\sum_{i=1}^{n_I}\log\left(p_s(s_i^{(I)})\right) - \frac{P_I}{n_E}\sum_{i=1}^{n_E}\log\left(p_s(s_i^{(E)})\right) \qquad (13)$$

$$H(s|\Omega_I;\mathbf{A}) = -\frac{1}{n_I}\sum_{i=1}^{n_I}\log\left(p_s^{(I)}(s_i^{(I)})\right), \qquad (14)$$

$$H(s|\Omega_E;\mathbf{A}) = -\frac{1}{n_E}\sum_{i=1}^{n_E}\log\left(p_s^{(E)}(s_i^{(E)})\right). \qquad (15)$$

And the approximation of effective information is

$$I(s;c|\mathbf{A}) = H(s|\mathbf{A}) - P(\Omega_I)H(s|\Omega_I;\mathbf{A}) - P(\Omega_E)H(s|\Omega_E;\mathbf{A}) \qquad (16)$$

Eq. (16) is the objective function that we are to maximize for the optimal solution.

3 Parameter Optimization on Grassmann Manifold

In this section, we introduce the algorithm to efficiently optimize the Maximum Effective Information Criteria.

3.1 Grassmann Manifold

3.1.1 Basic Concepts

¿From Eq. (16), we can see that the objective function is nonlinear and commonly there is no closed form solution. Moreover, the solution space of the objective function has the following two characteristics:

1. Orthogonality Constrained: As mentioned above, the matrix $\mathbf{A}$ characterizing a projection satisfies the orthogonality: $\mathbf{A}^T\mathbf{A} = \mathbf{I}$. Thus, the problem is an constrained optimization problem.

2. Rotation Invariance: Reviewing the calculation of metric value as in Eq.(2), we have that for any $p \times p$ orthogonal matrix $\mathbf{R}$,

$$s(\mathbf{u}|\mathbf{AR}) = \mathbf{u}^T\mathbf{ARR}^T\mathbf{A}^T\mathbf{u} = \mathbf{u}^T\mathbf{AA}^T\mathbf{u} = s(\mathbf{u}|\mathbf{A}). \qquad (17)$$

From the geometric view, multiplying an orthogonal matrix on the right is equivalent to rotating the column vectors of a matrix, thus $s(\mathbf{u}|\mathbf{A})$ is invariant w.r.t rotation.

Grassmann manifold offers an efficient mean to address the the optimization problem by exploiting the particular geometric properties of orthogonality and rotation invariance.

Considering that each subspace projection is characterized by an $n \times p$ matrix. From the view of differential geometry, in the $n \times p$-dimensional matrix space, all the $n \times p$ orthogonal matrices $\mathbf{A}$ satisfying $\mathbf{A}^T\mathbf{A} = \mathbf{I}_p$ constitute a continuous curved hypersurface in the np-dimensional space, called *Grassmann Manifold*, under the assumption that the objective function $F(\mathbf{A})$ defined upon it meets the *Homogeneity Condition* $F(\mathbf{A}) = F(\mathbf{AR})$, here $\mathbf{R}$ is a $p \times p$ orthogonal matrix satisfying $\mathbf{R}^T\mathbf{R} = \mathbf{RR}^T = \mathbf{I}$.

On the Grassmann manifold, if for two $n \times p$ orthogonal matrices $\mathbf{A}$ and $\mathbf{B}$, there exists a $p \times p$ orthogonal matrix $\mathbf{R}$ so that $\mathbf{B} = \mathbf{AR}$, then we call $\mathbf{A}$ and $\mathbf{B}$ *homogeneous*, denoted as $\mathbf{A} \sim \mathbf{B}$. It is obvious that the objective function F keeps its value unchanged for all matrices that are homogeneous, thus F can also be regarded as a function of equivalent classes of homogeneous matrices. An elaborate mathematical analysis of Grassmann manifold can be found in [1].

3.1.2 Tangent Space and Horizontal Space

As a curved hyper-surface, the movement of any point on the manifold always follows a direction on the *Tangent space*, which consists of all matrices $\mathbf{T}$ (the matrix space is a vector space, each matrix is a vector on it) that are tangent to the sub-manifold at the point. In mathematics, all the matrices in the tangent space at $\mathbf{A}$ satisfy $\mathbf{A}^T\mathbf{T} + \mathbf{T}^T\mathbf{A} = 0$ and for any matrix $\mathbf{Z}$, its projection on tangent space is [1]

$$\pi_T(\mathbf{Z}) = \frac{1}{2}\mathbf{A}(\mathbf{A}^T\mathbf{Z} - \mathbf{Z}^T\mathbf{A}) + (\mathbf{I} - \mathbf{AA}^T)\mathbf{Z}. \qquad (18)$$

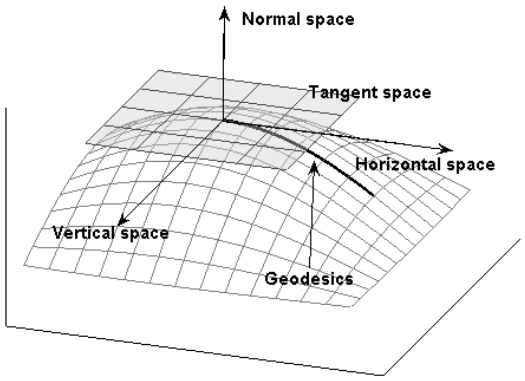

Figure 2. An illustration of Grassmann Manifold and related spaces. In optimization, the point should move on geodesics along the direction on the horizontal space. And the movement along direction on the vertical space will not change the value of objective function.

Considering the homogeneity condition, not all variations on the tangent space will result in the change of the objective function value. It can be shown that the tangent space can be orthogonally decomposed into the direct sum of a *vertical space* and a *horizontal space*, where only the directions in the horizontal space actually contribute to the change of objective function value. It is proved that the projection of any vector $\mathbf{Z}$ on the horizontal space at $\mathbf{A}$ is [1]

$$\pi_H(\mathbf{Z}) = (\mathbf{I} - \mathbf{A}\mathbf{A}^T)\mathbf{Z}. \tag{19}$$

3.1.3 Geodesics and Parallel Translation

In a flat plane, the shortest path from one point to another point is the straight line connecting these two points, and during the movement of a point, the tangent direction will not change. However, in a curved manifold, the situation is more complicated.

In differential geometry, the shortest path from one point to another point on the manifold is a curve, called *Geodesics*. On the Grassmann manifold, the geodesics going from the point $\mathbf{A}$ along the direction $\mathbf{H}$ can be represented by the following *Geodesic Equation* [1]:

$$\mathbf{A}(t) = [\mathbf{A}\mathbf{V}\cos(\mathbf{\Sigma}t) + \mathbf{U}\sin(\mathbf{\Sigma}t)]\mathbf{V}^T, \tag{20}$$

where $\mathbf{U}$, $\mathbf{\Sigma}$ and $\mathbf{V}$ can be obtained by performing Compact SVD on $\mathbf{A}$ as $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$. Here $\mathbf{U}$ is an $n \times p$ matrix, while $\mathbf{\Sigma}$ and $\mathbf{V}$ are $p \times p$ matrices.

Moreover, a tangent vector $\mathbf{\Delta}$ at that point will also be changed when moving along the geodesic, which can be computed by *Parallel Translation* as follows [1]:

$$\mathbf{\Delta}(t) = \left((-\mathbf{A}\mathbf{V}\sin(\mathbf{\Sigma}t) + \mathbf{U}\cos(\mathbf{\Sigma}t))\mathbf{U}^T + (\mathbf{I} - \mathbf{U}\mathbf{U}^T)\right)\mathbf{\Delta} \tag{21}$$

In the following description, the objective function is denoted as $F(\mathbf{A})$, while its gradient in the Euclidean space is denoted as $\nabla F(\mathbf{A})$

Step.1 Initialize with an $n \times p$ orthogonal matrix $\mathbf{A}^{(0)}$, and set $\mathbf{G}^{(0)} = (\mathbf{I} - \mathbf{A}^{(0)}\mathbf{A}^{(0)T})\nabla F(\mathbf{A}^{(0)})$, $\mathbf{H}^{(0)} = -\mathbf{G}^{(0)}$.

Step.2 Repeat the following steps until converged. At the k-th iteration ($k = 0, 1, \ldots$):

Step.2.1 Using linear search to find $t^{(k)}$ such that

$$t^{(k)} = \underset{t}{\operatorname{argmax}}\, F(\mathbf{A}^{(k)}(t))$$

Here, $\mathbf{A}^{(k)}(t)$ is computed as

$$\mathbf{A}^{(k)}(t) = [\mathbf{A}^{(k)}\mathbf{V}\cos(\mathbf{\Sigma}t) + \mathbf{U}\sin(\mathbf{\Sigma}t)]\mathbf{V}^T,$$

where $\mathbf{U}$, $\mathbf{\Sigma}$ and $\mathbf{V}$ are obtained by performing Compact SVD on $\mathbf{H}^{(k)}$.

Step.2.2 Update: $\mathbf{A}^{(k+1)} = \mathbf{A}^{(k)}(t^{(k)})$

Step.2.3 Compute $\mathbf{G}^{(k+1)}$ by projecting the gradient on horizontal space at $\mathbf{A}^{(k+1)}$

$$\mathbf{G}^{(k+1)} = (\mathbf{I} - \mathbf{A}^{(k+1)}\mathbf{A}^{(k+1)})\nabla F(\mathbf{A}^{(k+1)})$$

Step.2.4 Parallel transport the tangent vectors $\mathbf{H}^{(k)}$ and $\mathbf{G}^{(k)}$ as

$$\mathbf{H}^{(k)}(t^{(k)}) = (-\mathbf{A}^{(k)}\mathbf{V}\sin\mathbf{\Sigma}t^{(k)} + \mathbf{U}\cos\mathbf{\Sigma}t^{(k)})\mathbf{\Sigma}\mathbf{V}^T$$

$$\mathbf{G}^{(k)}(t^{(k)}) = \mathbf{G}^{(k)} - (\mathbf{A}^{(k)}\mathbf{V}\sin\Sigma t^{(k)} + \mathbf{U}(\mathbf{I} - \cos\mathbf{\Sigma}t^{(k)}))\mathbf{U}^T\mathbf{G}^{(k)}$$

Step.2.5 Update the tangent vector for next search as

$$\mathbf{H}^{(k+1)} = -\mathbf{G}^{(k+1)} + \gamma^{(k)}\mathbf{H}^{(k)}(t^{(k)})$$

$$\text{where} \qquad \gamma^{(k)} = \frac{\operatorname{tr}\left(\mathbf{G}^{(k+1)T}(\mathbf{G}^{(k+1)} - \mathbf{G}^{(k)}(t^{(k)}))\right)}{\operatorname{tr}\left(\mathbf{G}^{(k)T}\mathbf{G}^{(k)}\right)}$$

Step 2.6 Set $\mathbf{H}^{(k+1)} = -\mathbf{G}^{(k+1)}$ for every $p(n-p)$ iterations. The rationale is that the dimension of horizontal space is $p(n-p)$, thus there are $p(n-p)$ conjugate directions on it.

Table 1. Conjugate Gradient Optimization on Grassmann Manifold

The geometric explanation of a Grassmann manifold is illustrated in Figure 2.

3.2 Conjugate Gradient Optimization on Grassmann Manifold

By explicitly accounting for the geometric property embedded in the orthogonality constraint and rotation invariance, the original constrained optimization problem in the whole Euclidean space is converted to an unconstrained one on the Grassmann manifold.

Considering that in our maximum effective information criteria, it is difficult to compute the Hessian matrix, we adopt the conjugate gradient method for optimization and extend it to the Grassmann manifold with some critical changes: **1.** The gradient should be projected to be along the horizontal space at current submanifold as Eq. (19). **2.** In the 1-D searching step, the search is along a geodesic as Eq. (20) instead of a straight line. **3.** When the point is

moved, the tangent vector should be simultaneously parallel translated as Eq.(21). The whole procedure to maximize objective function using Conjugate Gradient Optimization on Grassmann manifold is briefly described in Table. 1.

3.3 Computation of Gradient

For the objective function of "Effective information", the gradient of the function is derived as follows:

$$\frac{\partial I(s;c|\mathbf{A})}{\partial \mathbf{A}} = \sum_{i=1}^{n} \frac{\partial H(s|\mathbf{A})}{\partial s_i} \frac{\partial s_i(\mathbf{u}_i|\mathbf{A})}{\partial \mathbf{A}}$$
$$-P(\Omega_I) \sum_{i=1}^{n_I} \frac{\partial H(s|\Omega_I;\mathbf{A})}{\partial s_i^{(I)}} \frac{\partial s_i^{(I)}(\mathbf{u}_i^{(I)}|\mathbf{A})}{\partial \mathbf{A}}$$
$$-P(\Omega_E) \sum_{i=1}^{n_E} \frac{\partial H(s|\Omega_E;\mathbf{A})}{\partial s_i^{(E)}} \frac{\partial s_i^{(E)}(\mathbf{u}_i^{(E)}|\mathbf{A})}{\partial \mathbf{A}}. \quad (22)$$

For a training set of n samples, we can acquire $\frac{1}{2}n(n-1)$ difference vectors, which is a huge quantity. Thus directly computing gradient is computational unaffordable. However, considering that the difference vectors available are far beyond sufficiency, it is enough to approximate the expectations with sufficient accuracy based on a set of subsampled metrics:

$$\frac{\partial H(s)}{\partial s_i} = \mathbf{f}(s_i; \{s_j\}|_{j=1}^{n}), \quad (23)$$
$$\frac{\partial H(s)}{\partial s_i} = \mathbf{f}(s_i^{(I)}; \{s_j^{(I)}\}|_{j=1}^{n_I}), \quad (24)$$
$$\frac{\partial H(s)}{\partial s_i} = \mathbf{f}(s_i^{(E)}; \{s_j^{(E)}\}|_{j=1}^{n_E}), \quad (25)$$
$$\frac{\partial s(\mathbf{u}|\mathbf{A})}{\partial \mathbf{A}} = \frac{\partial(\mathbf{u}\mathbf{A}^T\mathbf{A}\mathbf{u})}{\partial \mathbf{A}} = 2\mathbf{u}\mathbf{u}^T\mathbf{A}. \quad (26)$$

Here the function $\mathbf{f}$ is

$$\mathbf{f}(y;\{x_j\}|_{j=1}^{n}) = \frac{1}{nh}\left\{ -\frac{\sum_{k=1}^{n}\phi'(\frac{y-x_k}{h})}{\sum_{k=1}^{n}\phi(\frac{y-x_k}{h})} + \sum_{j=1}^{n}\frac{\phi'(\frac{x_j-y}{h})}{\sum_{k=1}^{n}\phi(\frac{x_j-x_k}{h})}\right\}. \quad (27)$$

$\{s_j^{(I)}\}|_{j=1}^{n_I}$ and $\{s_j^{(E)}\}|_{j=1}^{n_E}$ are obtained by subsampling the intrapersonal differences set and extrapersonal differences set respectively; while $\{s_i\}|_{j=1}^{n}$ is obtained by subsampling both sets with probability P_I and P_E.

4 Experiments

4.1 A Toy Problem

First of all, we consider a simple problem as depicted in Figure 3, where two classes of Gaussian distributed 2D samples are randomly generated. We would like to learn a projection direction, along which the samples from two classes can be best discriminated. Since the 1D projection is represented by a unit-length vector $\mathbf{w}$, thus the equation for Grassmann manifold is $\mathbf{w}^T\mathbf{w} = 1$, which is a circle on the 2D plane. The 3D curve in the figure represents the effective information on the manifold.

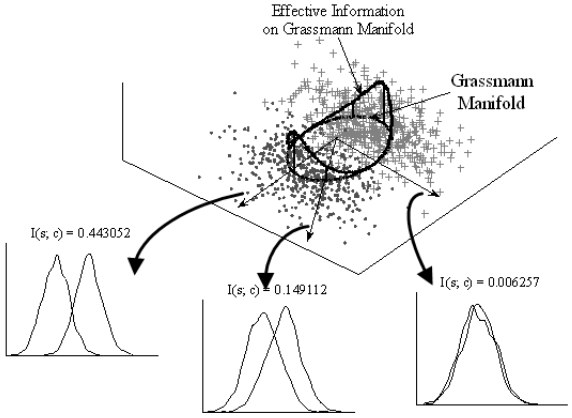

Figure 3. Illustration of Effective Information Principle in Toy Problem

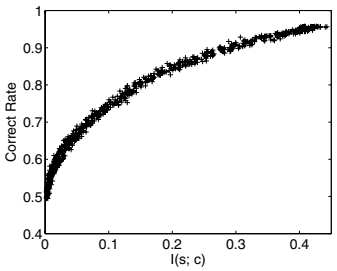

Figure 4. Effective Information vs. Correct Rate

¿From the toy example, we can have two observations: **(1) Effective information and classification is closely related.** We see that the effective information varies as the projection direction rotates along the manifold. When the direction is beneficial to discrimination, the effective information tends to be higher, which is shown more clearly in Figure 4, where we can see that the correct rate basically rises as the effective information increases. **(2)Effective information also reflects the degree of divergence between the distributions of intra-class difference and extra-class difference**. It can be seen in Figure 3 that, when effective information is high, the distributions are relatively far from the other and the overlap is small; when effective information approaches zero, the distributions are totally overlapped and the correct rate is near random guess.

4.2 Face Recognition

We test the proposed algorithm in the practical face recognition problem, which is a challenging arena to compare the effectiveness of different feature extraction methods. To clarify the following discussion, we first introduce the basic experiment procedure and how to evaluate the performance. There are two major tasks in face recognition: face classification and face verification. For both of the tasks, we divide all samples into three sets: training set, client set and probe set. Training set is employed to train the model. For our algorithm framework, the training process is simple. First we use LDA to estimate an initial projection matrix, then optimize it by maximizing the Effective Information using Conjugate Gradient on Grassmann Manifold. In testing, for face classification, each probe sample is compared with samples in the client set and classify the sample to the closest class in the projected subspace; while for face verification, each face is compared with samples in the client set and a judgement on whether the client face and probe face is from the same person is made based on the whether their distance in projected space is smaller than a threshold. The threshold value is adjusted so that the false reject rate and false accept rate on the ROC is the same. For both face classification and face identification, the performance is measured in terms of error rate, that is the ratio of the number of erroneous decisions to the number of all decisions made.

In our experiments, three databases are used for a thorough testing. The first is from FERET [11], where we select 995 samples from 298 persons for training with each person having 3 to 4 samples. 800 samples (fa) from 800 different persons comprise the client set, while the probe set consists of 800 samples (fb) from the same 800 persons captured in another session. The second is from XM2VTS [6], where there are 295 persons and each person has 4 samples captured in 4 different sessions. We use the first 3 sessions for training. The client samples are from the 1st session, while the probe samples are from the 4th session. The third is PURDUE [2], where there are 90 persons. For each person, we select 3 samples with various expressions for training, and select another 3 samples for testing. Among the 3 testing samples, one serves as the client sample, while the other two serve as probe samples.

Before training and testing, we first extract features from images to form vector representation as: (1) geometric normalization by cropping each image to a smaller image of size 64×72 pixels with eyes fixed at certain positions; (2) photometric normalization by histogram equalization and masking; (3) use appearance-based representation by taking the gray levels in order as vector components; and (4)use PCA to reduce dimension, where the dimension of principal subspace is determined by cross validation.

Database	Algorithm	Eff. Info.	Classific.	Verific.
FERET	PCA	0.2795	29.9%	9.56%
	LDA	0.4914	18.8%	6.40%
	ELDA	0.5989	8.50%	3.52%
	MEI	**0.7082**	**3.50%**	**2.85%**
XM2VTS	PCA	0.4626	27.5%	7.61%
	LDA	0.6291	9.49%	3.59%
	ELDA	0.6867	3.73%	3.08%
	MEI	**0.7571**	**2.71%**	**1.26%**
PURDUE	PCA	0.3286	19.4%	10.7%
	LDA	0.6388	7.22%	3.59%
	ELDA	0.7034	3.33%	2.57%
	MEI	**0.8102**	**1.67%**	**1.38%**

Table 2. Best Performance and Corresponding Effective Information Value.

In our experiments, we compare the performance of Maximum Effective Information (MEI) algorithm and other mainstream algorithms: PCA [7], LDA [10] and Enhanced LDA (ELDA) [3]. The comparative results in both face classification and face verification are illustrated in Figure 5. The detailed values of lowest error rates are listed in Table 2. To examine the relationship between effective information and performance, we also show the corresponding effective information value for the projection matrix trained by every algorithm. From the results, we see that our algorithm which maximizes the effective information consistently outperforms other state-of-the-art algorithms in face recognition literature for both face classification and face verification. Moreover, according to Table 2, the effective information has a close relationship with the performance, the higher effective information often follows by better accuracy.

For the experiments, it worths a discussion of the following points: **1.** Both the toy problem and face recognition experiments demonstrate the close relationship between effective information and classification performance, which validates the theoretical expectation. And the algorithm guided by MEI criteria also shows its remarkable superiority over other state-of-the-art algorithms. **2.** With assistance of Grassmann manifold, the freedom of optimization is drastically brought down and only 30 to 80 iterations are required to reach convergence.

5 Conclusion

In this paper, a novel channel model to interpret the classification process is formulated under the information theoretical perspective. Based on this model, we derive the maximum effective information principle and successfully tackle the computational obstacle. By employing Grass-

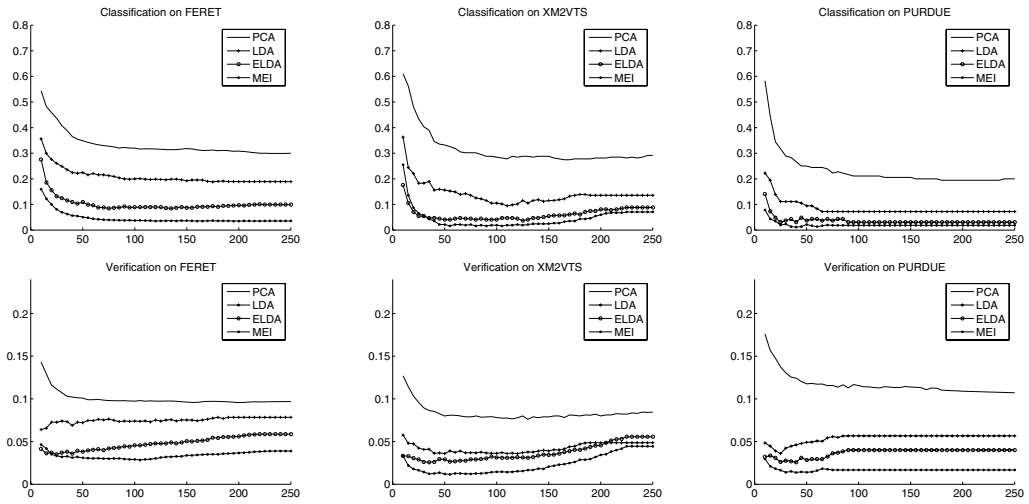

Figure 5. Comparative Results of algorithms on 3 databases: Performance vs. Number of features

mann manifold to exploit the geometric characteristics of solution space, the optimization procedure is significantly accelerated. The toy problem and extensive assessments in practical face recognition tasks convincingly show the significant improvement achieved by our algorithm.

Acknowledgement

The work described in this paper was fully supported by grants from the Research Grants Council of the Hong Kong Special Administrative Region. The work was conducted at the Chinese University of Hong Kong.

References

[1] A. Edelman, T.A. Arias, and S.T. Simth. The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Anal. Appl.*, 20(2):303–353, 2003.

[2] A. M. Martinez and R. Benavente. The ar face database. CVC Technical Report 24, Purdue Univ., 1998.

[3] C. Liu and H. Wechsler. Enhanced fisher linear discriminant models for face recognition. In *Proc. of CVPR'98*, 1998.

[4] C. Liu and H.Y. Shum. Kullback-leibler boosting. In *Proc. of CVPR'03*, 2003.

[5] R. Fano. Transmission of information: A statistical theory of communications. In *New York: Wiley*, 1961.

[6] J. Luettin and G. Maitre. Evaluation protocol for the extended m2vts database (xm2vts). Dmi for perceptual artificial intelligence, 1998.

[7] M. Turk and A. Pentland. Face recognition using eigenfaces. In *Proc. of CVPR'91*, 1991.

[8] M. Vidal-Naquet and S. Ullman. Object recognition with informative features and linear classification. In *Proc. of ICCV'03*, 2003.

[9] B. Moghaddam. Bayesian face recognition. *Pattern Recognition*, 33(11):1771–1782, 2000.

[10] P. Belhumeur, J. Hespanha, and D. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class-specific linear projection. *IEEE Trans. PAMI*, 19(7):711–720, 1997.

[11] P. J. Philips, H. Moon, S.A. Ryzvi, and P. J. Rauss. The feret evaluation methodology for face-recognition algorithms. *IEEE Trans. PAMI*, 12(10):1090–1104, 2000.

[12] R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern Classification (2nd Edition)*. Wiley Interscience Pub., 2001.

[13] K. Torkkola. Feature extraction by non-parametric mutual information maximization. *Journal of Machine Learning Research*, 3(7-8):1415–1438, 2003.

[14] N. Vasconcelos. Feature selection by maximum marginal diversity. In *Proc. of NIPS'02*, 2002.

[15] X. Wang and X. Tang. A unified framework for subspace face recognition. *IEEE Trans. PAMI*, 23(7):762–766, 2004.

[16] Y. Wu and A. Zhang. Feature selection for classifying high-dimensional numerical data. In *Proc of CVPR'04*, 2004.

A Appendix

Proof of Eq. (3)

From information theory, we have

$$\begin{aligned} I(x,y;z) &= H(z) - H(z|x,y) \\ &= H(z) - [H(z|x) - I(y;z|x)] \\ &= I(x;z) + I(y;z|x) \qquad \text{(a.0.1)} \end{aligned}$$

Because the y is uniquely and deterministically determined by x, i.e. $H(y|x) = H(y|x,z)$, we have

$$I(y;z|x) = H(y|x) - H(y|x,z) = 0. \qquad \text{(a.0.2)}$$

With (a.0.1) and (a.0.2), we have

$$\begin{aligned} I(x,y;z) &= I(x;z) = H(z) - H(z|x,y) \\ &= H(z) - [H(z|y) - I(x;z|y)] \\ &= I(y;z) + I(x;z|y). \qquad \text{(a.0.3)} \end{aligned}$$

Combine (a.0.3) and (a.0.3), the proposition in Eq. 3 is proved. □