METRICS OF GRASSMANNIAN REPRESENTATION IN REPRODUCING KERNEL HILBERT SPACE FOR VARIATIONAL PATTERN ANALYSIS

Yoshikazu WASHIZAWA

The University of Electro-Communications

ABSTRACT

Variation of patterns in signal can be represented by the covariance structure of vectors or its eigensubspace. When information of the pattern variation is available, representation by the covariance matrix or the eigensubspace is useful for feature extraction and classification compared with standard vector or matrix representations.

The structure and metric of the Grassmann manifold (Grassmannian) which is a set of eigensubspace, have been researched widely. Especially, the author has developed Mahalanobis distance in the Grassmannian, and it shows higher representation ability and classification accuracy compared with conventional Grassmannian representation methods.

In this paper, we extend Grassmannian metrics including the Mahalanobis distance using the kernel trick. We also propose an efficient basis vector selection algorithm and combine with the subset approximation of kernel principal component analysis to reduce the computational cost. In our experimental simulation by 3-dimensional object recognition problem, the proposed Mahalanobis distance shows better performance than conventional methods.

Index Terms— Subspace distance, Grassmann manifold, kernel trick, kernel principal component analysis, Mahalanobis distance

1. INTRODUCTION

In most of classical machine learning based signal processing methods, patterns are represented by vectors or matrices. Classical pattern analysis methods investigate the structure and metric of the matrices or a set of the vectors using the first and the second order statistics, such as covariance and eigensubspace. In this paper, we consider the case that each pattern is represented by its eigensubspace, and investigate the structure of the set of eigensubspaces which is called the Grassmann manifold.

For example, in face recognition problems, variations of angle, illumination, and facial expression will influence the pattern vector. The Grassmannian representation has been used to express such variations [1, 2]. Moreover, the Grassmann representation has been applied to handwritten character recognition, 3-D object recognition, and EEG classification problems [3, 4, 5].

A set of r-dimensional subspaces in d-dimensional input space is denoted by $\mathbb{G}(r,d)$. An element of $\mathbb{G}(r,d)$ is an r-dimensional subspace that has one-to-one correspondence with its orthogonal projection matrix $\boldsymbol{P}$. Thus we equate the subspace and its orthogonal projection matrix, $\boldsymbol{P} \in \mathbb{G}(r,d)$.

The structure and the distance metric of $\mathbb{G}(r,d)$ have been researched widely. Yamaguchi et al. proposed the mutual subspace method (MSM) [6]. The distance of MSM is defined as the minimum angle between two subspaces. The kernel version of MSM also has been proposed [7]. Hamm and Lee have proposed Grassmann discriminant analysis (GDA) which is a Grassmannian version of the linear discriminant analysis [8, 4]. GDA is also extended by using the kernel trick [1, 2]. The author proposed the Mahalanobis distance in Grassmannian [3, 5].

In this paper, we systematize the metrics in kernel Grassmannian, and propose the Mahalanobis distance in the reproducing kernel Hilbert space (RKHS). Unlike the standard extensions by kernel trick, since the metric operator of the Mahalanobis distance has to be expressed by basis vectors, its computational cost is not realistic if we simply use all available training vectors as the basis vectors. Therefore, we also propose an efficient basis selection algorithm. We compare the proposed method with several conventional methods by 3-D object classification problems. The proposed Mahalanobis distance showed the best classification performance.

2. GRASSMANNIAN REPRESENTATION

Let us consider a d-dimensional variational vector pattern $\boldsymbol{x}_0(\boldsymbol{t})$, where $\boldsymbol{t} = [t_1, \ldots, t_r]^\top$ denotes the variation factor such as the camera angle and illumination. Since it is difficult to directly deal with the function $\boldsymbol{x}_0(\boldsymbol{t})$, we approximate by using only the first order derivation,

$$\boldsymbol{x}_0(\boldsymbol{t}) \simeq \tilde{\boldsymbol{x}}_0(\boldsymbol{t}) = \boldsymbol{x}_0(\boldsymbol{0}) + \sum_{i=1}^{r} t_i \left(\frac{\partial \boldsymbol{x}_0(\boldsymbol{t})}{\partial t_i} \right) \Big|_{\boldsymbol{t}=\boldsymbol{0}}. \tag{1}$$

Let the Jacobian of $\boldsymbol{x}_0(\boldsymbol{t})$ be $\boldsymbol{J}_{\boldsymbol{x}_0} \in \mathbb{R}^{d \times r}$; $[\boldsymbol{J}_{\boldsymbol{x}_0}]_{ij} = \frac{\partial [\boldsymbol{x}_0(\boldsymbol{t})]_i}{\partial t_j}$. Then the pattern structure is represented by the linear manifold,

$$\mathcal{X} = \{\boldsymbol{x} | \boldsymbol{x} = \tilde{\boldsymbol{x}}_0(\boldsymbol{0}) + \boldsymbol{J}_{\boldsymbol{x}_0}|_{\boldsymbol{t}=\boldsymbol{0}} \boldsymbol{t},\ \boldsymbol{t} \in \mathbb{R}^r\}. \tag{2}$$

Alternatively, the minimum subspace that contains the linear manifold is considered, $\mathcal{X}' = \{\boldsymbol{x} | \boldsymbol{x} = \alpha \tilde{\boldsymbol{x}}_0(\boldsymbol{0}) + \boldsymbol{J}_{\boldsymbol{x}_0}|_{\boldsymbol{t}=\boldsymbol{0}} \boldsymbol{t},\ \alpha \in \mathbb{R}, \boldsymbol{t} \in \mathbb{R}^r\}$. If $\mathcal{X}$ contains the origin, $\mathcal{X} = \mathcal{X}'$, and if the input dimension d is sufficiently high, $\mathcal{X}$ and $\mathcal{X}'$ do not make a big difference.

The tangent distance method that has been proposed for character recognition problems employs the linear manifold representation, $\mathcal{X}$ [9]. On the other hand, Grassmann methods employ the subspace representation $\mathcal{X}'$.

In practical problems, the subspace $\mathcal{X}'$ is often obtained as the eigensubspace of a sequence of vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ such as moving image, continues shooting, and a camera array. Alternatively, the sequence can be generated from *a priori* knowledge of the variation. For example, in character recognition problems, we can artificially generate modified samples such as resizing, horizontal/vertical shift, rotation and so forth.

3. EXISTING METRICS ON GRASSMANNIAN

As we noted, a subspace $\mathcal{X}$ has one-to-one correspondence to the projection matrix onto the subspace, and thus we equate them. We introduce several existing distance measurements between two subspaces $\boldsymbol{P}_1, \boldsymbol{P}_2 \in \mathbb{G}(r, d)$.

3.1. Deterministic metrics

The projection metric $D_P(\boldsymbol{P}_1, \boldsymbol{P}_2)$ is simply defined by

$$D_{\mathrm{P}}(\boldsymbol{P}_1, \boldsymbol{P}_2) = \|\boldsymbol{P}_1 - \boldsymbol{P}_2\|_F, \tag{3}$$

where $\|\cdot\|_F$ denotes the Frobenius norm.

Let $\boldsymbol{U}_1, \boldsymbol{U}_2 \in \mathbb{R}^{d\times r}$ be matrices of the arbitrary orthonormal base of $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$ respectively. Then Binet-Cauchy metric D_{BC} is defined by

$$D_{\mathrm{BC}}(\boldsymbol{P}_1, \boldsymbol{P}_2) = 1 - \det(\boldsymbol{U}_1^\top \boldsymbol{U}_2). \tag{4}$$

Let σ_1 and σ_r be the maximum and minimum singular value of $\boldsymbol{U}_1^\top \boldsymbol{U}_2$ respectively. Then the minimum and maximum angles between $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$ are respectively $\theta_1 = \cos^{-1}\sigma_1$ and $\theta_r = \cos^{-1}\sigma_r$. The maximum and minimum correlations are defined by

$$D_{\max}(\boldsymbol{P}_1, \boldsymbol{P}_2) = \sin(\theta_1) \tag{5}$$

$$D_{\min}(\boldsymbol{P}_1, \boldsymbol{P}_2) = \sin(\theta_r). \tag{6}$$

It should be noted that $D_{\max}(\boldsymbol{P}_1, \boldsymbol{P}_2)$ does not satisfy the axiom of distances, i.e., $D_{\max}(\boldsymbol{P}_1, \boldsymbol{P}_2) = 0$ does not imply $\boldsymbol{P}_1 = \boldsymbol{P}_2$.

The metric of MSM is defined by the minimum angle θ_1 [6]. Since the sine function is monotonically increasing in $[-\pi/2, \pi/2]$, the maximum correlation and MSM are equivalent under the k-nearest neighbor rule.

These Grassmannian metrics are also defined by using principal angles [8]. GDA defines two kernels $K_{\mathrm{P}}(\boldsymbol{P}_1, \boldsymbol{P}_2) = \|\boldsymbol{U}_1^\top \boldsymbol{U}_2\|_F^2$ and $K_{\mathrm{BC}}(\boldsymbol{P}_1, \boldsymbol{P}_2) = (\det(\boldsymbol{U}_1^\top \boldsymbol{U}_2))^2$, and applies kernel LDA [8].

3.2. Mahalanobis distance on Grassmannian

In Euclidean space, the distance between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ under the positive definite metric matrix $\boldsymbol{M}$ is defined by $d_M(\boldsymbol{x}_1, \boldsymbol{x}_2) = (\boldsymbol{x}_1 - \boldsymbol{x}_2)^\top \boldsymbol{M}(\boldsymbol{x}_1 - \boldsymbol{x}_2)$. Let the covariance matrix of $\boldsymbol{x}$ be $\boldsymbol{\Sigma} = E[(\boldsymbol{x} - \bar{\boldsymbol{x}})(\boldsymbol{x} - \bar{\boldsymbol{x}})^\top]$, where $E[\cdot]$ is the ensemble mean, and $\bar{\boldsymbol{x}} = E[\boldsymbol{x}]$. The metric matrix of the Mahalanobis distance is defined by $\boldsymbol{M} = \boldsymbol{\Sigma}^{-1}$. This can be characterized by maximizing the following likelihood function,

$$L(\boldsymbol{M}|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{\det(\boldsymbol{M}^{-1})}} \exp\left(-\frac{1}{2} d_{\boldsymbol{M}}^2(\boldsymbol{x}_i, \bar{\boldsymbol{x}})\right). \tag{7}$$

L is maximized by $\boldsymbol{M} = \boldsymbol{\Sigma}^{-1}$ when $n \to \infty$.

In Grassmannian, the mean subspace of $\boldsymbol{P}_1, \ldots, \boldsymbol{P}_N$ under the projection metric is defined by [3]

$$\bar{\boldsymbol{P}} = \operatorname*{argmin}_{\boldsymbol{P}\in\mathbb{G}(r,d)} \frac{1}{N} \sum_{i=1}^{N} D_{\mathrm{P}}^2(\boldsymbol{P}, \boldsymbol{P}_i). \tag{8}$$

The mean subspace $\bar{\boldsymbol{P}}$ is given by the projection matrix onto the space spanned by r-major eigenvectors of $\sum_{i=1}^{N} \boldsymbol{P}_i$.

By introducing a positive definite metric matrix $\boldsymbol{M} \in \mathbb{R}^{d\times d}$, the projection metric is extended to

$$D_M(\boldsymbol{P}_1, \boldsymbol{P}_2) = \mathrm{Tr}[(\boldsymbol{P}_1 - \boldsymbol{P}_2)\boldsymbol{M}(\boldsymbol{P}_1 - \boldsymbol{P}_2)^\top]. \tag{9}$$

If $\boldsymbol{M}$ is the identity matrix $\boldsymbol{I}$, $D_M(\boldsymbol{P}_1, \boldsymbol{P}_2)$ is equivalent to the projection metric. In a similar manner to Eq. (7), we define the metric matrix by maximizing

$$L(\boldsymbol{M}|\boldsymbol{P}_1, \ldots, \boldsymbol{P}_N) = \prod_{i=1}^{N} \frac{1}{\sqrt{\det(\boldsymbol{M}^{-1})}} \exp\left(-\frac{1}{2} D_M^2(\boldsymbol{P}_i, \bar{\boldsymbol{P}})\right). \tag{10}$$

The solution is given by [3]

$$\boldsymbol{M} = \left(\frac{1}{N}\sum_{i=1}^{N} (\boldsymbol{P}_i - \bar{\boldsymbol{P}})(\boldsymbol{P}_i - \bar{\boldsymbol{P}})^\top\right)^{-1}. \tag{11}$$

In practice, we add a regularization term $\lambda\boldsymbol{I}$ before the inverse operation, where $\lambda \geq 0$ is the regularization parameter.

4. KERNEL EXTENSION OF GRASSMANNIAN METRICS

Several kernel extensions of Grassmannian metrics have been proposed based on kernel principal component analysis (KPCA) [1, 2, 7].

4.1. Kernel PCA and its subset approximation

In kernel trick, pattern vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are nonlinearly mapped to RKHS by $\phi(\cdot)$ which is derived by a positive definite kernel function, $k(\boldsymbol{x}_1, \boldsymbol{x}_2) = \langle\phi(\boldsymbol{x}_1), \phi(\boldsymbol{x}_2)\rangle$, $\forall \boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^d$, where $\langle\cdot,\cdot,\rangle$ is the inner product [10]. PCA obtains an orthonormal system $\boldsymbol{U} \in \mathbb{R}^{d\times r}$ that maximizes the variance, $\max_{\boldsymbol{U}|\boldsymbol{U}^\top\boldsymbol{U}=\boldsymbol{I}} \frac{1}{n}\sum_{i=1}^{n} \|\boldsymbol{U}^\top \boldsymbol{x}_i\|^2 = \boldsymbol{U}^\top \boldsymbol{R}\boldsymbol{U}$, where $\boldsymbol{R} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i^\top$ is the correlation matrix of $\boldsymbol{x}_i$.

In a similar way, KPCA obtains U by using mapped samples,

$$\max_{U|U^*U=I} \frac{1}{n}\sum_{i=1}^{n} \|U^*\phi(\boldsymbol{x}_i)\|^2, \tag{12}$$

where $\cdot^*$ denotes the adjoint operator that corresponds to the transpose in real finite vector space. Since $\phi(\boldsymbol{x}_i)$ could be an element of an infinite dimensional Hilbert space, we use the adjoint symbol and non-bold symbols. Let an operator,

$$\Phi = [\phi(\boldsymbol{x}_1), \ldots, \phi(\boldsymbol{x}_n)]. \tag{13}$$

Then since U is a bounded linear operator from the subspace spanned by $\phi(\boldsymbol{x}_1), \ldots, \phi(\boldsymbol{x}_n)$ to $\mathbb{R}^r$, U can be parameterized as $U = \Phi\boldsymbol{A}$, where $\boldsymbol{A} \in \mathbb{R}^{n\times r}$. The problem is reduced to $\max_{\boldsymbol{A}|\boldsymbol{A}^\top\boldsymbol{K}\boldsymbol{A}=\boldsymbol{I}} \mathrm{Tr}[\boldsymbol{A}^\top\boldsymbol{K}\boldsymbol{K}\boldsymbol{A}]$, where $\boldsymbol{K} = \Phi^*\Phi$ is the kernel Gram matrix ($[\boldsymbol{K}]_{i,j} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$). Suppose $\{\lambda_i, \boldsymbol{v}_i\}$ be the ith largest eigenvalue and corresponding eigenvectors of $\boldsymbol{K}$, and $\boldsymbol{V} = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_r] \in \mathbb{R}^{n\times r}$, $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_r)$. $\boldsymbol{A}$ is given by $\boldsymbol{A} = \boldsymbol{V}\boldsymbol{\Lambda}^{-1/2}$. For an input vector $\boldsymbol{x}$, the transform is $U^*\phi(\boldsymbol{x}) = \boldsymbol{A}^\top\boldsymbol{k}_{\boldsymbol{x}}$, where $\boldsymbol{k}_{\boldsymbol{x}} = [k(\boldsymbol{x}_1, \boldsymbol{x}), \ldots, k(\boldsymbol{x}_n, \boldsymbol{x})]^\top$.

KPCA is obtained by the eigenvalue decomposition (EVD) of the $n \times n$ matrix $\boldsymbol{K}$, and for an input vector, the transformed vector is obtained from n times evaluations of the kernel function. When n is large, these computational costs are high. When we extend

the Mahalanobis metric using the kernel trick, the projection matrix (operator) has to be expressed by the mapped basis vectors, $\phi(\boldsymbol{x}_1), \ldots, \phi(\boldsymbol{x}_n)$. In our framework, we deal with a number of patterns, and each pattern is expressed by n vectors. Therefore, if we use all available vectors, the computational complexity will be distant. In order to reduce the computational complexity of KPCA, we introduce the subset approximation (SubKPCA) [11].

SubKPCA uses the same objective function Eq. (12), but different base, that is U is parameterized $U = \Psi \boldsymbol{B}$, $\boldsymbol{B} \in \mathbb{R}^{m \times r}$, where $\Psi = [\phi(\boldsymbol{z}_1), \ldots, \phi(\boldsymbol{z}_m)]$. $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m$ are called the basis vectors, and we here suppose they are given. Then the problem is

$$\max_{\boldsymbol{B} | \boldsymbol{B}^\top \boldsymbol{K}_z \boldsymbol{B} = \boldsymbol{I}} \mathrm{Tr}[\boldsymbol{B}^\top \boldsymbol{K}_{xz}^\top \boldsymbol{K}_{xz} \boldsymbol{B}], \tag{14}$$

where $\boldsymbol{K}_z = \Psi^* \Psi$, ($[\boldsymbol{K}_z]_{i,j} = k(\boldsymbol{z}_i, \boldsymbol{z}_j)$), and $\boldsymbol{K}_{xz} = \Phi^* \Psi$, ($[\boldsymbol{K}_{xz}]_{i,j} = k(\boldsymbol{x}_i, \boldsymbol{z}_j)$). $\boldsymbol{B}$ is obtained as r-major eigenvectors of the generalized eigenvalue problem $\boldsymbol{K}_{xz}^\top \boldsymbol{K}_{xz} \boldsymbol{b} = \lambda \boldsymbol{K}_z \boldsymbol{b}$. In this case, the size of generalized EVD is m, and for an input vector, m times kernel evaluation is required. If $m = n$ and $\boldsymbol{x}_i = \boldsymbol{z}_i$ for $i = 1, \ldots, n$, SubKPCA is equivalent to KPCA, and m controls a trade-off between the approximation error and the computational complexity.

The selection of the basis vectors $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m$ is an important issue. However, in [11], simple sample selection methods such as random sampling consensus (RANSAC), clustering or forward/backward search were used to determine the basis vectors. We here propose an efficient sample selection based on the approximation error minimization in Section 5.1.

4.2. Deterministic metrics

For the ith pattern, we let an operator Φ_i and a matrix $\boldsymbol{A}_i$ as well as noted in Section 4.1. Then the projection operator onto r-dimensional subspace in RKHS is given by $P_i = \Phi_i \boldsymbol{A}_i \boldsymbol{A}_i^\top \Phi_i^*$.

The projection metric and Binet-Cauchy metric in RKHS is given by

$$\hat{D}_{\mathrm{P}}(P_1, P_2)^2 = \|P_1 - P_2\|_F^2 = 2r - 2\|\boldsymbol{A}_1^\top \boldsymbol{K}_{12} \boldsymbol{A}_2\|_F^2, \tag{15}$$

$$\hat{D}_{\mathrm{BC}}(P_1, P_2)^2 = 1 - \det(U_1^* U_2) = 1 - \det(\boldsymbol{A}_1^\top \boldsymbol{K}_{12} \boldsymbol{A}_2) \tag{16}$$

where $\boldsymbol{K}_{12} = \Phi_1^* \Phi_2$. The maximum and minimum correlations are also obtained by the maximum and minimum singular value of $\boldsymbol{A}_1^\top \boldsymbol{K}_{12} \boldsymbol{A}_2$ in a similar way. The kernel MSM was also proposed [7], and it is also equivalent to the kernel maximum correlation under the k-nearest rule. Kernel GDA uses the similar two kernels defined by the kernel Grassmann metrics, $\hat{K}_{\mathrm{P}}(P_1, P_2) = \|U_1^* U_2\|_F^2 = \|\boldsymbol{A}_1^\top \boldsymbol{K}_{12} \boldsymbol{A}_2\|_F^2$ and $\hat{K}_{\mathrm{BC}}(P_1, P_2) = \det(U_1^* U_2)^2 = \det(\boldsymbol{A}_1^\top \boldsymbol{K}_{12} \boldsymbol{A}_2)^2$.

5. PROPOSED METHOD

5.1. Basis selection method

Suppose that $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are all training vectors, $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m$ are basis vectors to be obtained, and $\Psi = [\phi(\boldsymbol{z}_1), \ldots, \phi(\boldsymbol{z}_m)]$. The ideal criterion to obtain the basis vectors is maximizing Eq. (12), where $U = \Psi \boldsymbol{B}$. However, it is difficult to optimize since the matrix $\boldsymbol{B}$ depends on Ψ. We thus i) iteratively obtain $\boldsymbol{z}_t$ for $t = 1, \ldots, m$; ii) assume $\{\boldsymbol{z}_t\}_{t=1}^m$ is the subset of all training vectors; and iii) approximate the criterion Eq. (12) by the full-rank case, i.e., projection onto the space spanned by current mapped basis vectors, In this case, we seek a vector whose projection norm onto the space is minimum,

$$\boldsymbol{z}_t = \operatorname*{argmin}_{\boldsymbol{z} \in \{\boldsymbol{x}_i\}_{i=1}^n} \frac{\|P_{\Psi_{t-1}} \phi(\boldsymbol{z})\|^2}{\|\phi(\boldsymbol{z})\|^2} = \operatorname*{argmin}_{\boldsymbol{z} \in \{\boldsymbol{x}_i\}_{i=1}^n} \frac{\boldsymbol{k}(\boldsymbol{z})^\top (\boldsymbol{K}_{t-1}^\Psi)^{-1} \boldsymbol{k}(\boldsymbol{z})}{k(\boldsymbol{z}, \boldsymbol{z})}, \tag{17}$$

where $P_{\Psi_{t-1}} = \Psi_{t-1}(\Psi_{t-1}^* \Psi_{t-1})^{-1} \Psi_{t-1}^*$ is the projection operator onto the subspace spanned by $\phi(\boldsymbol{z}_1), \ldots, \phi(\boldsymbol{z}_{t-1})$, $\boldsymbol{K}_{t-1}^\Psi = \Psi_{t-1}^* \Psi_{t-1}$, and $\boldsymbol{k}(\boldsymbol{z}) = \Psi_{t-1}^* \phi(\boldsymbol{z})$.

This process can be efficiently obtained by using updating of the Cholesky decomposition. Let $\boldsymbol{L}_{t-1}$ be the Cholesky decomposition of $\boldsymbol{K}_{t-1}^\Psi$ ($\boldsymbol{K}_{t-1}^\Psi = \boldsymbol{L}_{t-1} \boldsymbol{L}_{t-1}^\top$), and suppose we have $\boldsymbol{l}_{t-1}(\boldsymbol{x}_i) = \boldsymbol{L}_{t-1}^{-1} \boldsymbol{k}(\boldsymbol{x}_i)$ for all i. Then the numerator of Eq. (17) is $\|\boldsymbol{l}_{t-1}(\boldsymbol{x}_i)\|^2$, and for obtained $\boldsymbol{z}_t$, $\boldsymbol{l}_{t-1}(\boldsymbol{x}_i)$ can be updated by

$$\boldsymbol{l}_t(\boldsymbol{x}_i) = \begin{bmatrix} \boldsymbol{l}_{t-1}(\boldsymbol{x}_i) \\ \frac{1}{\sqrt{k(\boldsymbol{z}_t, \boldsymbol{z}_t) - \|\boldsymbol{l}_{t-1}(\boldsymbol{z}_t)\|^2}} (k(\boldsymbol{z}_t, \boldsymbol{z}_t) - (\boldsymbol{l}_{t-1}(\boldsymbol{z}_t))^\top \boldsymbol{l}_{t-1}(\boldsymbol{x}_i)) \end{bmatrix}. \tag{18}$$

We summarize the basis selection algorithm in Algorithm 1.

Algorithm 1 Basis selection algorithm

Require: Training vectors: $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, no. of basis vectors m, kernel function: $k(\cdot, \cdot)$
Ensure: Basis vectors: $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m$
1: Select the first basis vector $\boldsymbol{z}_1$
2: **for** $i = 1$ **to** n **do**
3: obtain initial $\boldsymbol{l}_t(\boldsymbol{x}_i) = k(\boldsymbol{z}_1, \boldsymbol{x}_i)$
4: **end for**
5: **for** $t = 2$ **to** m **do**
6: obtain $k = \mathrm{argmin}_i \|\boldsymbol{l}_{t-1}(\boldsymbol{x}_i)\|^2$, and set $\boldsymbol{z}_t = \boldsymbol{x}_k$.
7: Update $\boldsymbol{l}_t(\boldsymbol{x}_i)$ by Eq. (18) for all i
8: **end for**

5.2. Mahalanobis distance on kernel Grassmannian

We then obtain the mean subspace, that is defined by the projection operator onto the mean subspace. In a similar way to the case of the finite dimensional space, the projection operator is obtained from EVD of $\sum_{i=1}^N P_i$. We here use the SubKPCA, and each projection operator is given by $P_i = \Psi \boldsymbol{B}_i \boldsymbol{B}_i^\top \Psi^*$, where Ψ is common for all i, and obtained from the proposed basis selection algorithm. Let $\boldsymbol{L}_w$ is an arbitrary matrix that satisfies $\sum_{i=1}^N \boldsymbol{B}_i \boldsymbol{B}_i^\top = \boldsymbol{L}_w \boldsymbol{L}_w^\top$. Then eigenvectors of $\sum_{i=1}^N P_i$ is obtained from EVD of $\boldsymbol{L}_w^\top \boldsymbol{K}_z \boldsymbol{L}_w$. Let $(\lambda_i, \boldsymbol{v}_i)$ be the ith largest eigenvalue and corresponding eigenvector, and $\boldsymbol{V} = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_r] \mathrm{diag}(1/\sqrt{\lambda_1}, \ldots, 1/\sqrt{\lambda_r})$. Then the projector onto the mean subspace is given by $\bar{P} = \Psi \boldsymbol{V} \boldsymbol{V}^\top \Psi^*$.

The metric operator M of the Mahalanobis distance on kernel Grassmannian is given by

$$M = \Psi \boldsymbol{K}_z^{-1} (\frac{1}{N} \boldsymbol{Q})^{-1} \boldsymbol{K}_z^{-1} \Psi^* \tag{19}$$

$$\boldsymbol{Q} = \sum_{i=1}^N (\boldsymbol{B}_i \boldsymbol{B}_i^\top - \boldsymbol{V} \boldsymbol{V}^\top) \boldsymbol{K}_z (\boldsymbol{B}_i \boldsymbol{B}_i^\top - \boldsymbol{V} \boldsymbol{V}^\top), \tag{20}$$

We omit this detailed derivation since it is straightforward and space is limited.

We would add a regularization to the inverse operation of Eq. (19). From Eq. (11), the original inverse operation in the kernel Grassmannian is in RKHS, i.e., an inverse operation of $\Psi \boldsymbol{X} \Psi^*$ for an appropriate matrix $\boldsymbol{X}$. In this case, if we add the identity operator, i.e., $\Psi \boldsymbol{X} \Psi^* + \lambda I$, the inverse operation will be difficult because we also have to consider the complementary space. In order to avoid this, we add the projection operator onto the space spanned by Ψ, i.e., $\Psi \boldsymbol{X} \Psi^* + \lambda P_\Psi$. It is obvious if Ψ spans the whole space, $P_\Psi = I$. By changing the problem to finite dimensional space, the metric is given by,

$$M = \Psi \boldsymbol{K}_z^{-1} (\frac{1}{N} \boldsymbol{Q} + \lambda \boldsymbol{K}_z)^{-1} \boldsymbol{K}_z^{-1} \Psi^*. \quad (21)$$

The Mahalanobis distance between $P_1 = \Phi_1 \boldsymbol{A}_1 \boldsymbol{A}_1^\top \Phi_1^*$ and $P_2 = \Phi_2 \boldsymbol{A}_2 \boldsymbol{A}_2^\top \Phi_2^*$ is given by

$$\hat{D}_M(P_1, P_2) = \mathrm{Tr}[(P_1 - P_2) M (P_1 - P_2)]. \quad (22)$$

This can be calculated in finite dimensional matrix operations.

If we do not use the basis selection, but use all available training vectors, the size of the inverse operation in Eq. (19) and the size of EVD to obtain the mean subspace is nN, where N is the number of patterns, and each pattern has n vectors. This sometimes is too large to compute in practical computational environment.

6. EXPERIMENT

We used ETH-80 dataset [12]. The dataset consists of 3280 images of 3D objects (80 objects $\times$ 41 images of each from different angles). There are eight categories, 'apples,' 'pears,' 'tomatoes,' 'cows,' 'dogs,' 'horses,' 'cups,' and 'cars.' Each category contains ten objects. The task was to classify input patterns, each one containing the 41 images for a particular object. A subspace was obtained from the input images, and the distances between all template patterns and the input pattern were measured.

Each image has 128 $\times$ 128 pixels and has a mask image to remove its background. Each input pattern was preprocessed:

1. convert images to gray scale
2. remove background from each image using image masks
3. rescale images to 32 $\times$ 32 pixels using bicubic method
4. obtain histogram of gradients (HOG) features [13] for a dimension of 1296.

Each pattern thus had 41 vectors, and each vector had a dimension of 1296. The task was to classify each unlabeled pattern (with 41 vectors) into one of the eight classes.

We randomly selected 72 of the objects (90% of total) for use as template patterns; the remaining 8 objects (10%) were used as test patterns. We did this 100 times and generated 100 realizations. For each realization, we obtained the optimal hyper-parameters such as r by cross-validation using only the template patterns. Therefore, the estimated optimal hyper-parameters were different for each realization. The Gaussian kernel function $k(\boldsymbol{x}_1, \boldsymbol{x}_2) = \exp(-c\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|^2)$ was used, where c is obtained from $\{10^{-2}, 10^{-1.75}, \ldots, 10^{0}\}$ by the cross-validation. The regularization parameter λ is fixed to $\lambda = 0.1$. For GDA, we used the projection kernel $\hat{K}_{\mathrm{P}}(P_1, P_2) = \|U_1^* U_2\|_F^2$. We classified the patterns using the one nearest neighbor rule.

Table 1. Results of 3D object recognition experiment: mean test errors (Error) and standard deviations (SD). "-R" denotes RKHS, and "-I" denotes the input space.

Method	Error [%]	SD [%]
MAHAL-R	1.83	5.24
PROJ-R	2.50	6.43
BC-R	4.67	7.89
MSM-R	3.67	7.71
GDA-R	8.67	11.96
MAHAL-I	2.67	7.39
PROJ-I	6.50	9.45
BC-I	9.67	10.64
MSM-I	7.50	11.21
GDA-I	5.17	8.44

Table 1 lists the classification error rates and standard deviations. We compared five methods; i) the proposed Mahalanobis distance (MAHAL); ii) projection metric (PROJ); iii) Binet-Cauchy metric (BC); iv) MSM (or the maximum correlation); and v) GDA. "-R" denotes RKHS, and "-I" denotes the input space.

The proposed method, MAHAL-R exhibited the best performance. Except for GDA, kernel extensions improved classification accuracies.

7. CONCLUSION

We have proposed the Mahalanobis distance on kernel Grassmann distance. In order to reduce the computational complexity, we have also proposed a basis vector selection method. In our experiment by 3-D object recognition problem, the proposed method exhibited the best classification accuracy.

For future works, we apply the proposed method to various problems including time series data such as EEG data classification problems.

8. REFERENCES

[1] T. Wang and P. Shi, "Kernel Grassmannian distances and discriminant analysis for face recognition from image sets," *Pattern Recognition Letters*, vol. 30, no. 13, pp. 1161 – 1165, 2009.

[2] A. Yang and S. Chen, "An improved kernel discriminate analysis on Grassmannian manifold for face recognition," in *International Conference on Mechatronic Sciences, Electric Engineering and Computer (MEC)*, 2013, pp. 1000–1004.

[3] Y. Washizawa and S. Hotta, "Mahalanobis distance on grassmann manifold and its application to brain signal processing," in *Proc. of IEEE International Workshop on Machine Learning for Signal Processing (MLSP2012)*, 2012.

[4] J. Hamm and D.D. Lee, "Extended Grassmann kernels for subspace-based learning," in *Proc. of Neural Information Processing Systems*, 2008, pp. 601–608.

[5] Y. Washizawa and S. Hotta, "Mahalanobis distance on extended Grassmann manifolds for variational pattern analysis,"

IEEE Trans. on Neural Networks and Learning Systems (in press).

[6] O. Yamaguchi, K. Fukui, and K. Maeda, "Face recognition using temporal image sequence," in *Proc. of the 3rd International Conference on Face and Gesture Recognition*, 1998, p. 318.

[7] H. Sakano and N. Mukawa, "Kernel mutual subspace method for robust facial image recognition," in *Proc. of the 4th International Conference on Knowledge based Engineering System*, 2000, pp. 245–248.

[8] J. Hamm and D.D. Lee, "Grassmann discriminant analysis: a unifying view on subspace-based learning," in *Proc. of the 25th International Conference on Machine Learning (ICML2008)*, 2008, pp. 376–383.

[9] P.Y. Simard, Y. LeCun, J.S. Denker, and B. Victorri, "Transformation invariance in pattern recognition – tangent distance and tangent propagation," *International Journal of Imaging System and Technology*, vol. 11, no. 3, pp. 181–197, 2000.

[10] B. Schölkopf and A. Smola, *Learning with Kernels*, MIT press, 2002.

[11] Y. Washizawa, "Subset kernel principal component analysis," in *Proc. of 2009 IEEE International Workshop on Machine Learning for Signal Processing (MLSP 2009)*, 2009.

[12] B. Leibe and B. Schiele, "Analyzing appearance and contour based methods for object categorization," in *Proc. of Computer Vision and Pattern Recognition (CVPR 2003)*, 2003, pp. 409–415.

[13] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. of Computer Vision and Pattern Recognition (CVPR 2005)*, 2005, pp. 886–893.