Research Article

Hilbert–Schmidt Independence Criterion Subspace Learning on Hybrid Region Covariance Descriptor for Image Classification

Xi Liu, Peng Yang, Zengrong Zhan, and Zhengming Ma

1School of Information Engineering, Guangzhou Panyu Polytechnic, Guangzhou 511483, China
2School of Electronics and Information Technology, Sun Yat-Sen University, Guangzhou 510006, China

Correspondence should be addressed to Peng Yang; citystars@163.com

Received 10 December 2020; Revised 14 May 2021; Accepted 14 July 2021; Published 21 July 2021

1.Introduction

A growing number of non-Euclidean data, such as symmetric positive definite (SPD) manifolds [1] and Grassmann manifolds [2], are often encountered in vision recognition tasks. In particular, SPD manifolds have attracted increased attention in the form of the region covariance descriptor (RCD) [3, 4], Gaussian mixed model (GMM) [5], tensors [6–9], etc. In this work, we mainly discuss the image classification on SPD manifolds.

The RCD has been proved to be an effective descriptor in a variety of applications [10–12]. It captures the correlation between different features of an image and represents the image with a covariance matrix. However, the mean vector of features has been proved to be significant in image recognition tasks [13, 14]. In this work, we construct a new image descriptor by directly incorporating the mean feature information into the RCD. The new image descriptor is called the hybrid region covariance descriptor (HRCD). The HRCD inherits the advantages of the RCD, and it is more discriminable than the RCD. The images represented by the HRCD are also SPD matrices that lie on SPD manifolds. Most classical machine learning algorithms are constructed on linear spaces. Given the non-Euclidean geometry of Riemannian manifolds, directly using most of the conventional machine learning methods on Riemannian manifolds is inadequate [15, 16]. Therefore, the classification of the points on Riemannian manifolds has become a hot research topic.

Two main approaches are generally adopted to cope with the nonlinearity of Riemannian manifolds. The first approach is to construct learning methods by directly considering the Riemannian geometry; one such method is the widely used tangent approximation [17, 18]. Most existing SPD classification methods have been proposed by making use of Riemannian metrics [15, 16] or matrix divergences [19, 20] as the distance measure for SPD matrices [21–23].

The other approach is to project the SPD matrices to another space, such as a high-dimensional reproducing
kernel Hilbert space (RKHS) [24] and another low-dimensional SPD manifold [25]. Classification algorithms can be constructed on the projection space. Benefiting from the success of kernel methods in Euclidean spaces, the kernel-based classification scheme is a good choice for the analysis of SPD manifolds and has shown promising performance [26, 27]. Kernel-based methods embed manifolds into RKHSs and further project these manifolds to Euclidean spaces via an explicit mapping. Hence, algorithms designed for linear spaces can be extended to Riemannian manifolds. However, the mapping from RKHSs to Euclidean spaces using existing methods is based on a linear assumption. Moreover, the intrinsic connections of SPD matrices and low-dimensional projections are ignored.

To circumvent this limitation of kernel-based methods, we propose introducing the Hilbert–Schmidt independence criterion (HSIC) to the kernel trick and refer to the resulting method as the HSIC subspace learning (HSIC-SL) algorithm. Specifically, we derive the log-linear and log-Gaussian kernels to embed SPD matrices into a high-dimensional RKHS and then project these points into a low-dimensional vector space of the RKHS. To align the low-dimensional representation with the intrinsic features of the input data, we introduce statistical dependence between the SPD matrices and the low-dimensional representation. In this work, explicit mapping is obtained on the basis of subspace learning and HSIC maximization. Here, HSIC can be used to characterize the statistical correlation between two datasets.

The main contributions of this study are as follows:

1. We propose a novel covariance descriptor called the HRCD. The proposed descriptor explores discriminative information effectively.
2. The HSIC is first applied to the kernel framework on SPD Riemannian manifolds, and a novel subspace learning algorithm called the HSIC-SL is proposed. The proposed method achieves effective classification on the basis of global HSIC maximization.
3. We identify two simple kernel functions involved in the HSIC-SL algorithm. The diversity of kernels improves the flexibility of the HSIC-SL.

The rest of the paper is organized as follows. We provide a review of previous work in Section 2. A brief description about RCD, RKHS, and HSIC is presented in Section 3. We derive the proposed descriptor and algorithm in detail in Section 4. The experimental results are presented in Section 5 to demonstrate the effectiveness of the HRCD and HSIC-SL. Conclusions and future research directions are established in Section 6.

2. Literature Review

This section presents a brief review of RCDs, as well as recent manifold classification methods constructed on SPD manifolds.

The RCD was first introduced by Tuzel et al. [28]. It represents an image region with a nonsingular SPD matrix by extracting the covariance matrix of multiple features. The covariance matrix does not have any information about size and ordering, which implies certain scale and rotation independence. The RCD is used not only in image recognition but also in image set recognition tasks, in which an image set is modeled with its natural second-order statistic [4, 29]. The GMM could also serve as the SPD descriptor of an image set. Under the assumption of the multi-Gaussian distribution of an image set [30], hundreds of images in the image set are assigned to a small number of Gaussian components. Each Gaussian component is represented as an SPD matrix [31]. Thus, the image set is described by multiple SPD matrices. As mentioned previously, mean vectors have also been proved to be important in recognition tasks. In [32], the mean information was utilized in an improved log-Euclidean Gaussian kernel. However, this approach is limited to a specific algorithm and lacks generality. In the current work, we propose to incorporate the feature mean information and covariance matrix into a new SPD matrix and introduce first-order statistic information into the image RCD to improve the discriminant ability of the descriptor.

When the manifold under consideration is an SPD manifold, the tangent space of a particular point is a linear space. Most works map SPD matrices onto the tangent space of a particular point; thus, traditional linear classifiers can be applied. Under this framework, dimensionality reduction and clustering methods, such as Laplacian eigenmaps, local linear embedding (LLE), and Hessian LLE, have been extended to Riemannian manifolds [17]. Tuzel et al. introduced LogitBoost for classification on Riemannian manifolds [18]. The classifier has been generalized to multiclass classification [33]. Sparse coding by embedding manifolds into identity tangent spaces to identify the Lie algebra of SPD manifolds was considered in [34]. Such tangent space approximations could preserve manifold value data and eliminate the swelling effect. However, flattening a manifold through tangent spaces may generate inaccurate modeling, especially for regions far away from the tangent pole.

Except for tangent approximation, many efforts have been devoted to the distance measure on SPD manifolds to measure the true SPD manifold geometry; examples include the log-Euclidean Riemannian metric (LERM) [15] and the affine invariance Riemannian metric (AIM) [16]. Although matrix divergences are not real Riemannian metric, they provide fast and approximate distance computation. Sivalingam et al. proposed tensor sparse coding (TSC) for positive definite matrices [35] that utilizes the Burg divergence to perform sparse coding and dictionary learning on SPD manifolds. Riemannian dictionary learning and sparse coding (DLSC) [36] represents data as sparse combinations of SPD dictionary atoms via a Riemannian geometric approach and characterizes the loss of optimization for DLSC via the affine invariant Riemannian metric. However, these methods cannot be applied to other Riemannian manifolds because of the specificity of the specific metrics used. Embedding discriminant analysis (EDA) [37] identifies a bi-linear isometric mapping such that the resulting representation maximizes the preservation of Riemannian geodesic distance.
As for the proposed kernel methods for SPD manifolds, Riemannian locality preserving projections (RLPPs) [38] embed Riemannian manifolds into low-dimensional vector spaces by defining Riemannian kernels; moreover, their computational complexity is heavy, and the kernel is not always positive definite. Jayasumana et al. [39] presented a framework on Riemannian manifolds to identify the positive definiteness of Gaussian RBF kernels and utilized the log-Euclidean Gaussian kernel in kernel principal component analysis (KPCA) for a recognition task. Caseiro et al. proposed a heat kernel mean shift on Riemannian manifolds [40]. In [41], kernel DLSC based on LERM was introduced. Harandi et al. proposed to seek sparse coding by embedding the space of SPD matrices into Hilbert spaces through two types of Bregman matrix divergences [42]. Covariance discriminative learning (CDL) [4] utilizes a matrix logarithm operator to define kernel functions and then explicitly maps discriminative learning (CDL) [4] utilizes a matrix logarithm operator to define kernel functions and then explicitly maps.

Riemannian kernels. Considerable results are achieved in the Euclidean space. Zhuang et al. proposed a data-dependent operator to define kernel functions and then explicitly maps discriminative learning (CDL) [4] utilizes a matrix logarithm operator to define kernel functions and then explicitly maps.

Reproducing kernel Hilbert space (RKHS) is the theoretical basis of kernel methods. After projecting the data into a RKHS, various machine learning methods will be implemented in the RKHS.

Let \( \mathcal{S}(\Omega) \) be a function space, and \( \langle \cdot, \cdot \rangle \) is an inner product defined on \( \mathcal{S}(\Omega) \). The complete inner product space \( H = (\mathcal{S}(\Omega), \langle \cdot, \cdot \rangle) \) induced by \( \langle \cdot, \cdot \rangle \) is a Hilbert space. For all \( x \in \Omega \) and \( f \in \mathcal{S}(\Omega) \), if the function \( k \) satisfies \( f(x) = \langle f, k(x, \cdot) \rangle \), then \( k \) is the reproducing kernel of the RKHS \( H \). We denote the mapping defined by the reproducing kernel as \( \phi(x) = k(\cdot, x) = k_x \in H \). We can induce that

\[
\langle \phi(x), \phi(y) \rangle = \langle k_x(k_y, \cdot), y \rangle = k_x(y) = k(y, x) = k(x, y).
\]

The function \( k \) could be a kernel function only if the kernel matrix \( K \) is symmetric positive definite, where

\[
K = \begin{bmatrix}
 k(x_1, x_1) & \cdots & k(x_1, x_n) \\
 \vdots & \ddots & \vdots \\
 k(x_n, x_1) & \cdots & k(x_n, x_n)
\end{bmatrix}.
\]

According to Mercer’s theorem [45], once a valid reproducing kernel is defined, we can generate a unique Hilbert space.

3.3. Hilbert–Schmidt Independence Criterion (HSIC). The HSIC [46] is usually used to characterize the statistical correlation of two datasets. The mathematical theory of HSIC has been studied for a long time and there are many achievements [47–51]. In the computation of HSIC, the two datasets are firstly embedded onto two RKHSs, and then the HSIC of the two set of data is measured by the Hilbert–Schmidt (HS) operator of these two RKHSs.

Let \( X \) be a random variable/vector defined on \( \Omega_X \) and \( Y \) be a random variable/vector defined on \( \Omega_Y \). For two separate Hilbert spaces, and \( \phi_X: \Omega_X \rightarrow H_X \) and \( \phi_Y: \Omega_Y \rightarrow H_Y \) be the kernel mappings defined by the reproducing kernels, respectively.

3.3.1. Hilbert–Schmidt (HS) Operators. Let \( T: H_X \rightarrow H_Y \) be a compact operator and \( \{e_l^X\}_{l \in \mathbb{I}} \) be the orthonormal basis of \( H_X \); if \( \sum_{l \in \mathbb{I}} \|Te_l^X\|^2 < +\infty \), then \( T \) is called a Hilbert–Schmidt (HS) operator [52]. If for all \( T, S \in \text{HS}(H_X \rightarrow H_Y), \sum_{l \in \mathbb{I}} |\langle Te_l^X, Se_l^Y \rangle| < +\infty \), then \( \text{HS}(H_X \rightarrow H_Y), \langle \cdot, \cdot \rangle_{\text{HS}} \) is a Hilbert space. The inner product \( \langle \cdot, \cdot \rangle_{\text{HS}} \) is defined as \( \langle T, S \rangle_{\text{HS}} = \sum_{l \in \mathbb{I}} \langle Te_l^X, Se_l^Y \rangle \). For \( f_0 \in H_X \) and \( g_0 \in H_Y \), the tensor product of \( f_0 \) and \( g_0 \) is denoted as \( f_0 \otimes g_0 \). Since \( f_0 \otimes g_0 \) is in \( H_Y \), then \( f_0 \otimes g_0 \in \text{HS}(H_X \rightarrow H_Y) \) [53].

3.3.2. Mean Functions and Cross Covariance Operators. Let \( \Phi_X: H_X \rightarrow R \) be a continuous linear functional over \( H_X \); for all \( T \in \text{HS}(H_X \rightarrow H_Y), \Phi_X(f) = \langle f, \mu_X \rangle_{\chi^X} \); then, \( \mu_X \) is called the mean function of \( \Phi_X(f) \). Similarly, the mean function \( \mu_Y \) of \( \Phi_Y(Y) \) is defined in the same way.

3. Related Work

In this section, we briefly review the RCD and the properties of the RKHS and HSIC.

3.1. Region Covariance Descriptor. Region covariance descriptor (RCD), as a special case of SPD matrices, proposes a natural way of fusing multiple features. Suppose \( R \) is an image region of size \( h \times w \), and we can extract multiple features of every pixel in \( R \). The features could be location, grey values, and gradients. We denote the feature vector of the \( k \)-th pixel as

\[
z_k = \left[ x, y, I_x, I_x^T, I_y, I_y^T \right],
\]

where \( x \) and \( y \) denote the location, \( I \) is the grey value, and \( I_x \) and \( I_y \) are the gradients with respect to \( x \) and \( y \). The RCD of \( R \) is defined as

\[
\Sigma = \frac{1}{n - 1} \sum_{k=1}^{n} (z_k - \mu)(z_k - \mu)^T,
\]

where \( n = h \times w \) and \( \mu = (1/n) \sum_{k=1}^{n} z_k \in R^d \) denotes the mean of the points. Then, the image region can be presented by a \( d \times d \) SPD matrix, where \( d \) depends on the number of features.

3.2. RKHS. Reproducing kernel Hilbert space (RKHS) is the theoretical basis of kernel methods. After projecting the data into a RKHS, various machine learning methods will be implemented in the RKHS.

Let \( \mathcal{S}(\Omega) \) be a function space, and \( \langle \cdot, \cdot \rangle \) is an inner product defined on \( \mathcal{S}(\Omega) \). The complete inner product space \( H = (\mathcal{S}(\Omega), \langle \cdot, \cdot \rangle) \) induced by \( \langle \cdot, \cdot \rangle \) is a Hilbert space. For all \( x \in \Omega \) and \( f \in \mathcal{S}(\Omega) \), if the function \( k \) satisfies \( f(x) = \langle f, k(x, \cdot) \rangle \), then \( k \) is the reproducing kernel of the RKHS \( H \). We denote the mapping defined by the reproducing kernel as \( \phi(x) = k(\cdot, x) = k_x \in H \). We can induce that

\[
\langle \phi(x), \phi(y) \rangle = \langle k_x(k_y, \cdot), y \rangle = k_x(y) = k(y, x) = k(x, y).
\]

The function \( k \) could be a kernel function only if the kernel matrix \( K \) is symmetric positive definite, where

\[
K = \begin{bmatrix}
 k(x_1, x_1) & \cdots & k(x_1, x_n) \\
 \vdots & \ddots & \vdots \\
 k(x_n, x_1) & \cdots & k(x_n, x_n)
\end{bmatrix}.
\]

According to Mercer’s theorem [45], once a valid reproducing kernel is defined, we can generate a unique Hilbert space.
Let $\Phi$ be a continuous linear functional over $\text{HS}(H_X \rightarrow H_Y)$; for all $T \in \text{HS}(H_X \rightarrow H_Y)$,

$$\Phi(T) = E_{XY} \left[ \langle \varphi_X(X) \otimes \varphi_Y(Y), T \rangle_{\text{HS}} \right].$$

(5)

Then, according to Riesz theorem, there must be a unique HS operator $C_{XY} \in \text{HS}(H_X \rightarrow H_Y)$, such that for all $T \in \text{HS}(H_X \rightarrow H_Y)$,

$$\Phi(T) = \langle T, C_{XY} \rangle_{\text{HS}},$$

(6)

where $C_{XY}$ is the cross covariance operator between $\varphi_X(X)$ and $\varphi_Y(Y)$.

The relationship between $C_{XY}, \mu_X$, and $\mu_Y$ is illustrated in Figure 1. The two datasets $\Omega_X$ and $\Omega_Y$ are embedded into $H_X$ and $H_Y$ by the kernel functions $\varphi_X: \Omega_X \rightarrow H_X$ and $\varphi_Y: \Omega_Y \rightarrow H_Y$, respectively. $\mu_X$ and $\mu_Y$ are the mean functions. The HSIC of $\Omega_X$ and $\Omega_Y$ is given by the Hilbert–Schmidt (HS) operator $C_{XY}$ of $H_X$ and $H_Y$.

![Figure 1: The sketch mapping of HSIC.](image)

3.3.3. HSIC. HSIC of two random variables/vectors is defined as follows:

$$\text{HSIC}(X,Y) = E_{XY} \left[ \left\| (\varphi_X(X) - \mu_X) \otimes (\varphi_Y(Y) - \mu_Y) \right\|_{\text{HS}}^2 \right].$$

(7)

It can be seen from the definition of $\text{HSIC}(X,Y)$ that instead of directly calculating the covariance of $X$ and $Y$, i.e., $E_{XY} \left[ (X - E_X[X])(Y - E_Y[Y]) \right]$, HSIC first transforms $X$ and $Y$ into $H_X$ and $H_Y$, respectively, and then calculates the covariance of $\varphi_X(X)$ and $\varphi_Y(Y)$ by using HS operators between $H_X$ and $H_Y$. In practice, $H_X$ and $H_Y$ are generated from kernel functions $k_X$ and $k_Y$.

If the joint probability distribution of $X$ and $Y$ is given or known, $\text{HSIC}(X,Y)$ can be calculated as follows:

$$\text{HSIC}(X,Y) = E_{XY} \left[ \left\| (\varphi_X(X) - \mu_X) \otimes (\varphi_Y(Y) - \mu_Y) \right\|_{\text{HS}}^2 \right] = \langle C_{XY}, C_{XY} \rangle_{\text{HS}} - 2\langle C_{XY}, \mu_X \otimes \mu_Y \rangle_{\text{HS}} + \langle \mu_X \otimes \mu_Y, \mu_X \otimes \mu_Y \rangle_{\text{HS}}.$$

(8)

where

$$\langle C_{XY}, C_{XY} \rangle_{\text{HS}} = E_{XY} E_{X'Y'} \left[ k_X(X, X') k_Y(Y, Y') \right],$$

(9)

$$\langle C_{XY}, \mu_X \otimes \mu_Y \rangle_{\text{HS}} = E_X [E_Y [k_X(X, X') k_Y(Y, Y')]],$$

(10)

$$\langle \mu_X \otimes \mu_Y, \mu_X \otimes \mu_Y \rangle_{\text{HS}} = \langle \mu_X, \mu_X \rangle_X \langle \mu_Y, \mu_Y \rangle_Y.$$
with all elements being 1 and gives $H$ we embed the SPD matrices into a high-dimensional RKHS. Each training sample is described by an SPD matrix. Second, the kernel function $k$ of the feature vectors of each point in $R$ and then compute the mean vector and covariance matrix of the features. Suppose that the feature vector of the $k$-th pixel is $z_k$; the mean vector $\mu \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$ can then be computed as

$$\mu_i = \frac{1}{n} \sum_{k=1}^{n} z_k \in \mathbb{R}^d,$$

$$\Sigma = \frac{1}{n-1} \sum_{k=1}^{n} (z_k - \mu)(z_k - \mu)^T.$$  

Following the information geometry theory [54], we combine the mean and the covariance matrix into a new matrix without additional computational complexity. The new matrix is constructed as

$$X = [\Sigma^{-(1/(d+1))}] + \frac{\mu \mu^T}{d+1}.$$  

Here, $d$ is the dimensionality of the feature vector, and $|.|$ is the determinant operator. The $(d+1) \times (d+1)$ SPQ matrix $X$ is the HRCD of the image. As a result of the inheritance from the covariance matrix, the HRCD is not only effective, robust, and low-dimensional but also more discriminable than the RCD.

### 4.2. Kernel Function in HSIC-SL

In defining a valid RKHS, the kernel must be symmetric positive definite. Many discussions on the symmetric positive definiteness of kernel functions are based on vector spaces. In this section, we introduce two typical kernel functions on SPD Riemannian manifolds.

#### 4.2.1. Log-Linear Kernel

The polynomial kernel is one of the commonly used kernel functions on Euclidean spaces. The polynomial kernel function in a vector space is defined as

$$k(x_i, x_j) = (\alpha x_i^T x_j + \beta)^d,$$

where $x_i, x_j \in \mathbb{R}^d$. If $\alpha = \gamma = 1$ and $\beta = 0$, then equation (17) is a linear kernel.

$$k(x_i, x_j) = x_i^T x_j.$$  

When the linear kernel is developed into SPD Riemannian manifolds, it should be redefined in a sophisticated form. The linear kernel on SPD Riemannian manifolds can be defined as

$$k(x_i, x_j) = x_i^T x_j.$$
4.2. Log-Gaussian Kernel. The Gaussian kernel is another popular kernel function in Euclidean spaces. The definition of a Gaussian kernel function is

$$k(x_i, x_j) = \exp\left(-\frac{d^2(x_i, x_j)}{2\sigma^2}\right), \quad \text{(20)}$$

where $d(x_i, x_j) = \|x_i - x_j\|_p$. A good effect can be achieved by replacing the Euclidean distance with a log-Euclidean distance. The log-Gaussian kernel is defined by

$$k_{\text{LE}}(x_i, x_j) = \exp\left\{-\frac{d_{\text{LE}}^2(x_i, x_j)}{2\sigma^2}\right\}, \quad \text{(21)}$$

where $d_{\text{LE}}(x_i, x_j) = \|\log(x_i) - \log(x_j)\|_p$ is the log-Euclidean distance between $x_i$ and $x_j$. The positive definiteness of $k_{\text{LE}}$ was proved in [39].

The parameter $\sigma$ is an important parameter in the Gaussian kernel. To make the log-Gaussian kernel sensitive to distances, we suggest setting $\sigma$ to the average value of distances between the training samples.

4.3. HSIC Subspace Learning. After embedding the matrices to the RKHS, we further project the points into a vector space through explicit mapping. We aim to find the explicit mapping from the RKHS to the vector space by maximizing the HSIC between the SPD matrices and the low-dimensional representation, as well as preserving the local information. The proposed HSIC-SL includes global HSIC maximization and within-class information preservation.

We denote the HSIC of $\chi$ and the low-dimensional representation $Y$ as $\text{HSIC}(\chi, Y)$. According to equation (14), $\text{HSIC}(\chi, Y)$ can be computed as

$$\text{HSIC}(\chi, Y) = \frac{1}{N^2} \text{tr}(K_{\chi}C_NK_{\chi}C_N). \quad \text{(22)}$$

The input data $\chi$ and projection $Y$ are represented by $K_{\chi}$ and $K_{Y}$, respectively. To explicitly realize the low-dimensional representation, we define the kernel function of $Y$ in $\text{HSIC}(\chi, Y)$ as $k_{Y}: R^d \times R^d \rightarrow R$, $\forall y', y'' \in R^d$; that is,

$$k_{Y}(y', y'') = y'^T y'' \quad \text{(23)}$$

We denote the kernel matrix of $k_{Y}$ as $K_{Y}$. It can be computed by

$$K_{Y} = \begin{bmatrix} y_1^T y_1 & \cdots & y_1^T y_N \\ \vdots & \ddots & \vdots \\ y_N^T y_1 & \cdots & y_N^T y_N \end{bmatrix} = Y^T Y \quad \text{(24)}$$

Substituting equation (24) to equation (22) yields

$$\text{HSIC}(\chi, Y) = \frac{1}{N^2} \text{tr}(K_{\chi}C_NK_{\chi}C_N) = \frac{1}{N^2} \text{tr}(Y^T Y C_N K_{\chi} C_N) \quad \text{(25)}$$

As $N$ is not related to $Y$, the coefficient $(1/N^2)$ in equation (25) can be omitted. Then, we have

$$\text{HSIC}(\chi, Y) = \text{tr}(W K_{\chi} C_N K_{\chi} C_N K_{\chi}^TW^T) = \text{tr}(L_{H} W^T), \quad \text{(26)}$$

where $L_{H} = K_{\chi} C_N K_{\chi} C_N K_{\chi}^T$.

The within-class information is represented by the within-class scatter $S_{W}$, which is defined as

$$S_{W} = \text{tr}\left(\sum_{i=1}^{c} \frac{N_i}{N} (y_k - m_i)(y_k - m_i)^T\right), \quad \text{(27)}$$

where $N_i$ is the number of training samples of the $i$-th class, $\sum_{i=1}^{c} N_i = N$, and $m_i$ is the mean vector corresponding to the
$\text{Equation (27)}$ can be further transformed into

$$S_W = \text{tr} \left( W \left( \sum_{i=1}^{c} \sum_{j=1}^{N_i} (K_{j \text{Row}} - K_m)(K_{j \text{Row}} - K_m)^T \right) W^T \right) = \text{tr} \left( WL_W W^T \right),$$

where $K_m = (1/N_i) \sum_{j=1}^{N_i} k_j^T$, $L_W = \sum_{i=1}^{c} \sum_{j=1}^{N_i} (K_{j \text{Row}} - m_i)^T (K_{j \text{Row}} - m_i)^T$.

In sum, the objective function is formulated as

$$J(W) = \arg\max_{W} \left( \text{HSIC}(x, y) \right) = \arg\max_{W} \left( WL_H W^T \right)$$

s.t. $\text{tr}(WL_W W^T) = \text{tr}(L_W).$

(29)

The Rayleigh quotient maximum problem is commonly used in optimization problems because of the fast and simple calculation. The problem shown in equation (29) can be solved by calculating the Rayleigh quotient maximum. To tackle the singularity, we add a small perturbation $\epsilon$ to the diagonal elements of $L_W$. The optimal projection matrix $W$ is composed of the eigenvectors corresponding to the $m$ biggest eigenvalues of $(L_W + \epsilon I_N)^{-1} L_H$, where $I_N$ is the identity matrix.

Hence, for the given test image, we first compute its HRCD and denote the result as $X_t$. The projection can be obtained by $y_t = W K_{i \text{Row}} \text{row}$. Then, the class of the test image can be predicted through the nearest neighbor classifier.

5. Experiment

The performance of the HRCD and the proposed algorithm is verified in this section. We considered five widely studied image datasets: COIL-20 (Columbia Object Image Library) dataset [55], ETH-80 dataset [56], Queen Mary University of London (QMUL) dataset [57], face data FERET dataset [58], and Brodatz dataset [59]. All of the compared methods were implemented in MATLAB R2014 and tested on an Intel(R) Core(TM) i5-4670K (3.40GHz) machine.

5.1. Performance of HRCD. To verify that the HRCD is an effective image descriptor, we directly used the KNN classifier on the image feature space represented by the HRCD and RCD without feature extraction. By adopting the Euclidean metric, LERM, AIRM, and Burg divergence as the measurements, the classification experiments were performed on COIL-20 and ETH-80. The COIL-20 dataset contains 20 objects, each of which contains 72 images measuring $128 \times 128$ at different directions. Figure 3 shows the sample pictures. Features including grey values and first- and second-order gradients were extracted to calculate the RCD and HRCD of an image. Hence, the RCD and HRCD of an image were a $5 \times 5$ SPD matrix and a $6 \times 6$ SPD matrix, respectively. The images were randomly split into the training set and test set, with 10 pictures assigned to the training set and the remaining images assigned to the test set.

ETH-80 is an image set containing eight types of objects, such as apple, pears, cars, and dogs. Each object has 10 instances, and each instance contains images from 41 different viewpoints. The images in ETH-80 were resized to $128 \times 128$ (Figure 4). For the RCD and HRCD representations, we extracted the following features:

$$F(x, y) = \left[ x, y, R_{x,y}, G_{x,y}, B_{x,y}, I_{x,y}, I_x, I_y, I_{xx}, I_{yy} \right].$$

(30)

where $R_{x,y}, G_{x,y}, B_{x,y}$ are the RGB color values of a pixel at the position of $x$ and $y$, $I_{x,y}$ is the greyscale value, and $I_x, I_y, I_{xx}, I_{yy}$ are the first- and second-order gradients of intensities. The RCD and HRCD of the image were a $10 \times 10$ SPD matrix and a $11 \times 11$ SPD matrix, respectively. Half of the instances in every object were used for training, and the remaining instances were used for the test. Each instance in the training and test sets comprised 100 random samples. Therefore, the training and test sets each contained 800 images.

Table 1 lists the classification accuracies and runtimes under different metrics. To eliminate the randomness of the experiment, we obtained the average accuracy and runtime for 20 tests.

5.2. Performance of HSIC-SL. The proposed HSIC-SL was compared with several recognition methods on SPD manifolds. The compared methods included RLPP [38], KSLR [32], CDL [4], KPCA using the log-Gaussian kernel [39], RSR [42], TSC [35], Riem-DLSC [36], logEu-c-SC [34], Geometry-DR [25], KLKR-DL [43], EDA [37], and MKSSCR [44]. For brevity, we denote the HSIC-SL with the log-linear kernel as HSIC-SL (log-linear) and that with the log-Gaussian kernel as HSIC-SL (log-Gaussian). HSIC-SL (log-linear) and HSIC-SL (log-Gaussian) were combined with the RCD and HRCD. Thus, for the proposed HSIC-SL, four different combinations were tested. For equality, the important parameters of the comparison methods were set according to the suggestion of the original paper.

5.2.1. Experiments on QMUL Dataset. The QMUL dataset [44] is a set of images of human heads collected from airport terminal cameras. The dataset is composed of 20,005 images. It is divided into five classes according to the direction of the head images: back, front, left, right, and background. The samples from QMUL are shown in Figure 5. The dataset was divided into the training and test sets in advance. Table 2 shows the number of training and test sets in every class. The extracted feature of any pixel is
where $I_L(x, y)$, $I_a(x, y)$, and $I_b(x, y)$ are the three channel values of the CIELAB color space, $I_x$ and $I_y$ are the first-order gradients in the $x$- and $y$-directions of $I_L(x, y)$, respectively, and $G_i(x, y); i = 1, \ldots, 8$ is the response of eight difference-of-Gaussians filters. We obtained a $13 \times 13$ SPD matrix for the RCD and a $14 \times 14$ SPD matrix for the HRCD. The training data consisted of 200 randomly selected samples for each category, and the test set consisted of 100 randomly selected samples. The KNN ($k = 12$) search was used to construct the neighborhood graphs in the RLPP and Geometry-DR. The parameters ($\sigma$) in the kernels of the KPCA, RLPP, KSLR, and HSIC-SL were set to the average distances. The parameter $\gamma$ in the KSLR was set to 0.3. The parameter $\epsilon$ in the proposed method was set to 0.001. We evaluated the performance of the CDL, RLPP, KSLR, Geometry-DR, and HSIC-SL for various dimensions and reported the maximum performance. In logEuc-SC, RSR, TSC, Riem-DLSC, and KLRM-DL, 50 dictionaries and kernel parameters were learned from the training set. The kernel function in the RSR and the basic kernel in the KLRM-DL was the Stein kernel. The parameter $alpha$ was set to 0.1, and the number of data samples was set to 30. The 1NN classifier was adopted in all the algorithms.

In Table 3, we show the recognition accuracy of the HSIC-SL and the other existing algorithms. To eliminate the
randomness of the experiment, we used the average recognition rate for 20 tests. HSIC-SL (log-Gaussian) + HRCD and HSIC-SL (log-linear) + HRCD achieved impressive performance while HSIC-SL (log-Gaussian) + RCD obtained the highest classification accuracy. Moreover, the accuracy of the HRCD was greater than that of the RCD in the experiment. These results indicated that the HRCD was better than the RCD. Furthermore, HSIC-SL + HRCD was better than the other algorithms.

5.2.2. Experiments on FERET Dataset. To conduct the face recognition experiment, we used the "b" subset of the FERET dataset [56], which consists of 2,000 face images of 200 people. The images are those of 71 females and 129 males of diverse ethnicities, genders, and ages. The images were cropped and downsampled to $64 \times 64$. The training set was composed of images with "ba," "bc," "bh," and "bk" labels. Images marked as "bd," "be," "bf," and "bg" constituted the test set. The feature vector for computing the RCD and HRCD is described by

$$F(x, y) = [x, y, I(x, y), G_{00}(x, y), \ldots, G_{47}(x, y)],$$

(32)

where $x$ and $y$ denote the position, $I(x, y)$ is the intensity, and $G_{uv}(x, y)$ is the response value of the Gabor filter. The direction $u$ of the Gabor filter was from 0 to 4, and the scale $v$ was from 0 to 7. Thus, the RCD and HRCD of each image were a $43 \times 43$ SPD matrix and a $44 \times 44$ SPD matrix, respectively. The neighborhood graphs constructed in the RLPP and Geometry-DR were KNN ($k = 3$). The kernel functions with Jeffrey and Stein divergences were adopted in RSR and, respectively, denoted as RSR-J and RSR-S for brevity. In RSR, TSC, Riem-DLSC, KLRM-DL, and logEuc-SC, all training samples were regarded as dictionary atoms. The settings of the other parameters were the same as those for the QMUL dataset.

Table 4 shows the recognition rates of the compared algorithms. The proposed method was not the best algorithm for the FERET dataset. It only achieved the highest recognition accuracy in the "bd" test scenario. Nevertheless, the average recognition accuracies of HSIC-SL were still better than those of the other algorithms and were only slightly worse than those of KLRLM-DL. Hence, HSIC-SL was still a feasible algorithm for the FERET dataset. We also noticed that HSIC-SL (log-Gaussian) performed better than HSIC-SL (log-linear). Therefore, the log-Gaussian kernel was more suitable than the log-linear kernel for this dataset.

5.2.3. Experiments on Brodatz Dataset. We performed two texture classification experiments on the Brodatz dataset [57]. Examples from the Brodatz dataset are shown in Figure 5: Sample images of QMUL dataset.

| Table 1: Comparison of RCD and HRCD in terms of classification accuracy (%) and runtime (seconds). |
| --- | --- | --- | --- |
| Metric | Descriptor | COIL-20 Classification accuracy (%) | Runtime (s) | ETH-80 Classification accuracy (%) | Runtime (s) |
| Euclidean | RCD | 74.98 | 2.946 | 62.63 | 5.234 |
| | HRCD | 59.88 | 2.908 | 66.34 | 5.299 |
| LERM | RCD | 84.81 | 4.149 | 71.03 | 7.1 |
| | HRCD | 88.99 | 4.262 | 72.04 | 7.759 |
| AIRM | RCD | 87.10 | 118.33 | 71.64 | 482.47 |
| | HRCD | 91.06 | 125.79 | 73.35 | 509.15 |
| Burg divergence | RCD | 89.23 | 10.102 | 72.07 | 26.368 |
| | HRCD | 91.71 | 10.492 | 73.63 | 26.66 |
5.2.4. Experiments on COIL-20 and ETH-80 Datasets. In this experiment, we used the COIL-20 dataset [55] and ETH-80 dataset in the object categorization task. The experimental procedure was the same as that described in Section 5.1. We compared the proposed method with KPCA [39], RLPP [38], KSLR [32], and CDL [4]. In addition, KPCA and RLPP were conducted on the HRCD and, respectively, denoted as KPCA + HRCD and RLPP + HRCD. The classifier adopted in all of the algorithms was the 1NN classifier.

Table 7 shows the classification accuracies of the methods on COIL-20 and ETH-80. First, HSIC-SL obtained the best accuracy in all of the datasets. This result indicated that the introduction of the HSIC improved the effectiveness of the recognition algorithm. Second, the classification accuracies of RLPP, KPCA, and HSIC-SL in the RCD were lower than those in the HRCD (i.e., RLPP + HRCD, KPCA + HRCD, and HSIC-SL + HRCD). This result proved once again that the HRCD had advantages over the RCD. Finally, the effectiveness of the log-linear kernel and log-Gaussian kernel in HSIC-SL was demonstrated in the experiments.

5.3. Analysis of Dimensionality. The parameter $m$ was regarded as the dimensionality of the vector space after feature extraction. The curves of the classification accuracies of the compared algorithms on COIL-20 [55], ETH-80 [37], and Brodatz versus $m$ are shown in Figures 8 and 9. The experimental setups were the same as those described in the previous section.

With the increase of the dimensionality, the recognition accuracy curves showed an upward trend. When the recognition accuracy reached a certain value, the recognition rate remained basically stable within a certain range of the subspace dimension.

5.4. Discussion. In the above experiments, the performance of the RCD and HRCD and the effectiveness of HSIC-SL and the other algorithms were compared. The following observations were made:

(1) The classification accuracy in the image feature space represented by the HRCD was better than that by the RCD regardless of which classifier was used (i.e., KNN classifier without feature extraction or the proposed HSIC-SL). The result showed that the proposed image descriptor HRCD outperformed the RCD.

(2) When the RCD was used as the image descriptor, the HSIC-SL method was superior to most of the methods, except for the FERET and Brodatz datasets. In FERET, the performance of Riem-DLSC, MKSSCR, and KLRM-DL was slightly better than that of HSIC-SL + RCD (log-Gaussian kernel). In Brodatz, the performance of HSIC-SL + RCD was slightly worse than that of the other methods in the 5-texture group, 10-texture group, and 16-texture group. Nevertheless, the recognition accuracy of HSIC-SL + RCD in the experiment on all texture images was higher than those of the other methods. The results showed that HSIC-SL was indeed an excellent algorithm on SPD manifolds, but it was...
inferior in the classification of datasets with subtle features, such as face recognition and texture recognition. At the same time, the HRCD makes up for this defect to a certain extent. The performance of HSIC-SL + HRCD was almost superior to that of all methods. However, in the FERET dataset, the average recognition accuracy of HSIC-SL was lower than that of KLRM-DL.
| Methods                              | COIL-20 | ETH-80 |
|-------------------------------------|---------|--------|
| KPCA                                | 81.05   | 72.61  |
| KPCA + HRCD                         | 83.79   | 73.7   |
| RLPP                                | 85.89   | 74.08  |
| RLPP + HRCD                         | 88.79   | 75.63  |
| CDL                                 | 94.54   | 79.92  |
| KSLR                                | 96.24   | 81.66  |
| HSIC-SL (log-linear) + RCD          | 96.72   | 82.80  |
| HSIC-SL (log-linear) + HRCD         | 97.75   | 84.60  |
| HSIC-SL (log-Gaussian) + RCD        | 96.87   | 82.40  |
| HSIC-SL (log-Gaussian) + HRCD       | 97.92   | 85.28  |

**Figure 8:** Recognition rates versus different dimensionalities on COIL-20 database.

**Figure 9:** Recognition rates versus different dimensionalities on ETH-80 database.
(3) In the experiments, we also compared the performance of the log-Gaussian kernel and log-linear kernel. In general, the log-Gaussian kernel was better than the log-linear kernel. However, in the experiments on QMUL and Brodatz, the log-linear kernel obtained better results than the log-Gaussian kernel. The difference in performance indicated that the choice of kernel affected the performance of HSIC-SL. We can improve the performance of HSIC-SL by selecting a suitable kernel function.

6. Conclusions

In this work, we propose an improved covariance descriptor called the HRCD, which represents images with SPD matrices. The HRCD inherits the advantages of the RCD and is more effective.

To address the classification problem on SPD Riemannian manifolds, we propose an efficient image classification method that is based on a kernel framework. We refer to it as HSIC-SL. Through the definition of the log-linear kernel and log-Gaussian kernel, the input images represented by SPD matrices can be embedded into the RKHS. To seek explicit mapping from the RKHS to the vector space, HSIC-SL constructs the objective function on the basis of the framework of subspace learning and HSIC maximization. HSIC-SL always outperforms other representative methods without increasing computational complexity.

The proposed algorithm also has certain limitations. The average classification accuracy is slightly worse than that of KLRM-DL on the FERET dataset. Hence, the covariance descriptor is not strong enough to handle the classification of small details, such as face recognition. For our future work, we will employ other effective features to form the covariance matrices. We will also explore other useful kernel functions to suit different types of datasets.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This study was supported by the National Natural Science Foundation of China through the Project “Research on Nonlinear Alignment Algorithm of Local Coordinates in Manifold Learning” under grant no. 61773022, the Character and Innovation Project of Education Department of Guangdong Province under grant no. 2018GKTSCX081, the Young Innovative Talents Project of Education Department of Guangdong Province under grant no. 2020QKNCX191, the Guangzhou Science and Technology Plan Project of Bureau of Science and Technology of Guangzhou Municipality under grant no. 202102020700, and the Educational Big Data Enterprise Lab of Guangzhou Panyu Polytechnic under grant no. 2021XQS05.

References

[1] V. Arsigny, P. Fillard, X. Pennec, and N. Ayache, “Geometric means in a novel vector space structure on symmetric positive-definite matrices,” SIAM Journal on Matrix Analysis and Applications, vol. 29, no. 1, pp. 328–347, 2007.

[2] M. Harandi, C. Sanderson, C. Shen, and B. Lovell, “Dictionary learning and sparse coding on Grassmann manifolds: an extrinsic solution,” in Proceedings of the 2013 IEEE International Conference on Computer Vision, pp. 3120–3127, Sydney, Australia, 2013.

[3] F. Porikli, O. Tuzel, and P. Meer, “Covariance tracking using model update based on Lie algebra,” in Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, New York, NY, USA, 2006.

[4] R. Wang, H. Guo, L. S. Davis, and Q. Dai, “Covariance discriminative learning: a natural and efficient approach to image set classification,” in Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 2012.

[5] W. Wang, R. Wang, Z. Huang, S. Shan, and X. Chen, “Discriminant analysis on Riemannian manifold of Gaussian distributions for face recognition with image sets,” IEEE Transactions on Image Processing, vol. 27, no. 1, pp. 151–163, 2017.

[6] I. L. Dryden, A. Kolloidenko, and D. Zhou, “Non-Euclidean statistics for covariance matrices, with applications to diffusion tensor imaging,” The Annals of Applied Statistics, vol. 3, no. 3, pp. 1102–1123, 2009.

[7] D. Le Bihan, J.-F. o. Mangin, C. Poupon et al., “Diffusion tensor imaging: concepts and applications,” Journal of Magnetic Resonance Imaging, vol. 13, no. 4, pp. 534–546, 2001.

[8] R. Caseiro, P. Martins, J. F. Henriques, and J. Batista, “A nonparametric Riemannian framework on tensor field with application to foreground segmentation,” Pattern Recognition, vol. 45, no. 11, pp. 3997–4017, 2012.

[9] M. G. Thomason and J. Gregor, “Higher order singular value decomposition of tensors for fusion of registered images,” Journal of Electronic Imaging, vol. 20, no. 1, pp. 13–23, 2011.

[10] Z. Gao, Y. Wu, M. Harandi, and Y. Jia, “A robust distance measure for similarity-based classification on the SPD manifold,” IEEE Transactions on Neural Networks and Learning Systems, vol. 31, no. 9, pp. 3230–3244, 2020.

[11] Y. Wu, Y. Jia, P. Li, J. Zhang, and J. Yuan, “Manifold kernel sparse representation of symmetric positive-definite matrices and its applications,” IEEE Transactions on Image Processing, vol. 24, no. 11, pp. 3729–3741, 2015.

[12] A. Sanin, C. Sanderson, M. T. Harandi, and B. C. Lovell, “Spatio-temporal covariance descriptors for action and gesture recognition,” in Proceedings of the 2013 IEEE Workshop on Applications of Computer Vision (WACV), Clearwater Beach, FL, USA, 2013.

[13] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek, “Image classification with the Fisher vector: theory and practice,” International Journal of Computer Vision, vol. 105, no. 3, pp. 222–245, 2013.

[14] H. Jegou, F. Perronnin, M. Douze, J. Sanchez, P. Perez, and C. Schmid, “Aggregating local image descriptors into compact codes,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 9, pp. 1704–1716, 2012.
[15] X. Pennec, P. Fillard, and N. Ayache, “A Riemannian framework for tensor computing,” *International Journal of Computer Vision*, vol. 66, no. 1, pp. 41–66, 2006.

[16] V. Arsigny, P. Fillard, X. Pennec, and N. Ayache, “Log-Euclidean metrics for fast and simple calculus on diffusion tensors,” *Magnetic Resonance in Medicine*, vol. 56, no. 2, pp. 411–421, 2006.

[17] A. Goh and R. Vidal, “Clustering and dimensionality reduction on Riemannian manifolds,” in *Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, AK, USA, 2008.

[18] O. Tuzel, F. Porikli, and P. Meer, “Pedestrian detection via classification on Riemannian manifolds,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 10, pp. 1713–1727, 2008.

[19] B. Kulis, M. Sustik, and I. Dhillon, “Learning low-rank kernel matrices,” in *Proceedings of the 23rd International Conference on Machine Learning*, Pittsburgh, PA, USA, 2006.

[20] S. Sra, “A new metric on the manifold of kernel matrices with application to matrix geometric means,” in *Proceedings of the 25th International Conference on Neural Information Processing Systems*, vol. 1, Lake Tahoe, NV, USA, 2012.

[21] D. Tosato, M. Farenzena, M. Spera, V. Murino, and M. Cristani, “Multi-class classification on Riemannian manifolds for video surveillance,” in *Proceedings of the 2010 ECCV*, Heraklion, Greece, 2010.

[22] J. Carreira, R. Caseiro, J. Batista, and C. Sminchisescu, “Semantic segmentation with second-order pooling,” in *Proceedings of the 2012 ECCV*, Florence, Italy, 2012.

[23] A. Cherian, S. Sra, A. Banerjee, and N. Papanikolopoulos, “Jensen-Bregman LogDet divergence with application to efficient similarity search for covariance matrices,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 9, pp. 2161–2174, 2013.

[24] N. Aronszajn, “Theory of reproducing kernels,” *Transactions of the American Mathematical Society*, vol. 68, no. 3, pp. 337–404, 1950.

[25] M. Harandi, M. Salzmann, and R. Hartley, “Dimensionality reduction on SPD manifolds: the emergence of geometry-aware methods,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 1, pp. 48–62, 2018.

[26] M. T. Harandi, C. Sanderson, R. Hartley, and B. C. Lovell, “Sparse coding and dictionary learning for symmetric positive definite matrices: a kernel approach,” in *Proceedings of the ECCV 2012*, Florence, Italy, 2012.

[27] R. Vemulapalli, J. K. Pillai, and R. Chellappa, “Kernel learning for extrinsic classification of manifold features,” in *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition*, Portland, OR, USA, 2013.

[28] O. Tuzel, F. Porikli, and P. Meer, “Region covariance: a fast descriptor for detection and classification,” in *Proceedings of the ECCV 2006*, vol. 3952, Graz, Austria, 2006.

[29] H. Tan and Y. Gao, “Patch-based principal covariance discriminative learning for image set classification,” *IEEE Access*, vol. 5, pp. 15001–15012, 2017.

[30] O. Arandjelovic, G. Shakhnarovich, J. Fisher, R. Cipolla, and T. Darrell, “Face recognition with image sets using manifold density divergence,” in *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05)*, San Diego, CA, USA, 2005.

[31] M. Lovrić, M. Min-Oo, and E. A. Ruh, “Multivariate normal distributions parametrized as a Riemannian symmetric space,” *Journal of Multivariate Analysis*, vol. 74, no. 1, pp. 36–48, 2000.

[32] X. Liu and Z. Ma, “Kernel-based subspace learning on Riemannian manifolds for visual recognition,” *Neural Processing Letters*, vol. 51, no. 1, pp. 147–165, 2020.

[33] R. Caseiro, P. Martins, J. F. Henriques, F. S. Leite, and J. Batista, “Rolling Riemannian manifolds to solve the multi-class classification problem,” in *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition*, vol. 9, Portland, OR, USA, 2013.

[34] K. Guo, P. Ishwar, and J. Konrad, “Action recognition from video using feature covariance matrices,” *IEEE Transactions on Image Processing*, vol. 22, no. 6, pp. 2479–2494, 2013.

[35] R. Sivalingam, D. Boley, V. Morellas, and N. Papanikolopoulos, “Tensor sparse coding for positive definite matrices,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 3, pp. 592–605, 2014.

[36] A. Cherian and S. Sra, “Riemannian dictionary learning and sparse coding for positive definite matrices,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 12, pp. 2859–2871, 2017.

[37] X. Xie, Z. L. Yu, Z. Gu, and Y. Li, “Classification of symmetric positive definite matrices based on bilinear isometric Riemannian embedding,” *Pattern Recognition*, vol. 87, pp. 94–105, 2019.

[38] M. T. Harandi, C. Sanderson, A. Willems, and B. Lovell, “Kernel analysis over Riemannian manifolds for visual recognition of actions, pedestrians and textures,” in *Proceedings of the 2012 IEEE Workshop on the Applications of Computer Vision (WACV)*, Breckenridge, CO, USA, 2012.

[39] S. Jayasumana, R. Hartley, M. Salzmann, H. Li, and M. Harandi, “Kernel methods on Riemannian manifolds with Gaussian RBF kernels,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 12, pp. 2464–2477, 2015.

[40] R. Caseiro, J. F. Henriques, P. Martins, and J. Batista, “Semi-intrinsic mean shift on Riemannian manifolds,” in *Proceedings of the ECCV 2012*, vol. 7572, Florence, Italy, 2012.

[41] P. Li, Q. Wang, W. Zuo, and L. Zhang, “Log-Euclidean kernels for sparse representation and dictionary learning,” in *Proceedings of the 2013 IEEE International Conference on Computer Vision*, Sydney, Australia, 2013.

[42] M. T. Harandi, R. Hartley, B. Lovell, and C. Sanderson, “Sparse coding on symmetric positive definite manifolds using Bregman divergences,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 6, pp. 1294–1306, 2016.

[43] R. Zhuang, Z. Ma, W. Feng, and Y. Lin, “SPD data dictionary learning based on kernel learning and Riemannian metric,” *IEEE Access*, vol. 8, pp. 61956–61972, 2020.

[44] S. Hechmi, A. Gallas, and E. Zagrouba, “Multi-kernel sparse learning on Riemannian manifolds for video surveillance,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, pp. 2859–2871, 2017.

[45] J. Shawetaylor and N. Cristianini, *Kernel Methods for Pattern Analysis*, Cambridge University Press, New York, NY, USA, 2004.

[46] E. Kreyszig, *Introductory Functional Analysis with Applications*, Vol. 1, Wiley, New York, NY, USA, 1978.

[47] B. B. Damodaran, N. Courty, and S. Lefèvre, “Sparse Hilbert Schmidt independence criterion and surrogate-kernel-based feature selection for hyperspectral image classification,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 4, pp. 2385–2398, 2017.

[48] K. Zhang, V. W. Zheng, Q. Wang, J. T. Kwok, Q. Yang, and I. Marsic, “Covariate shift in Hilbert space: a solution via surrogate kernels,” in *Proceedings of the 30th International Conference on Machine Learning*, 2017.
M. J. Gangeh, H. Zarkoob, and A. Ghodsi, “Fast and scalable feature selection for gene expression data using Hilbert-Schmidt independence criterion,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 14, no. 1, pp. 167–181, 2017.

L. Song, A. Gretton, K. M. Borgwardt, and A. J. Smola, “Colored maximum variance unfolding,” in *Proceedings of the 20th International Conference on Neural Information Processing Systems*, Vancouver, Canada, 2007.

L. Song, A. Smola, A. Gretton, J. Bedo, and K. Borgwardt, “Feature selection via dependence maximization,” *Journal of Machine Learning Research*, vol. 13, no. 1, pp. 1393–1434, 2007.

J.-P. Aubin, “Hilbert-Schmidt operators and tensor products,” in *Applied Functional Analysis*, pp. 283–308, John Wiley & Sons, Hoboken, NJ, USA, 2nd edition, 2000.

A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf, "Measuring statistical dependence with Hilbert-Schmidt norms," in *Proceedings of the 2005 International Conference on Algorithmic Learning Theory*, pp. 63–77, Singapore, 2005.

S. I. Amari and H. Nagaoka, *Methods of Information Geometry*, Oxford University Press, New York, NY, USA, 2000.

S. Nene, S. Nayar, and H. Murase, “Columbia object image library (coil-20),” Technical report, Columbia University, New York, NY, USA, 1996.

B. Leibe and B. Schiele, “Analyzing appearance and contour based methods for object categorization,” in *Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, Madison, WI, USA, 2003.

D. Tosato, M. Spera, M. Cristani, and V. Murino, “Characterizing humans on Riemannian manifolds,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1972–1984, 2013.

P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, “The FERET evaluation methodology for face-recognition algorithms,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 10, pp. 1090–1104, 2000.

T. Randen and J. H. Husoy, “Filtering for texture classification: a comparative study,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 4, pp. 291–310, 1999.